

THE VICTORIAN PANEL SURVEY

**A scoping study for the ESRC
October 2003**

Contact:

*K. Schürer
UK Data Archive
University of Essex
Wivenhoe Park
Colchester
CO4 3SQ*

Tel: 01206 872008

Email: schurer@essex.ac.uk

CONTENTS

1 EXECUTIVE SUMMARY	4
2 BACKGROUND	5
2.1 RATIONALE	5
2.2 THE VPS IN OUTLINE	5
3 RESEARCH ISSUES	8
3.1 POTENTIAL USE OF A VPS	8
3.2 THE EVIDENCE OF MODERN LONGITUDINAL STUDIES	8
3.3 POTENTIAL APPLICATIONS FOR THE VPS	9
4 SOURCES.....	14
4.1 HISTORICAL CENSUS DATA.....	14
4.2 HISTORICAL REGISTRATION DATA	16
4.3 THE 1851 CENSUS SAMPLE.....	19
5 METHODOLOGY UTILISED IN EXISTING LONGITUDINAL STUDIES ...	22
5.1 THE BRITISH HOUSEHOLD PANEL SURVEY.....	22
5.2 THE LONGITUDINAL STUDY.....	24
5.3 OTHER UK COHORT STUDIES.....	25
5.3.1 <i>The National Child Development Survey.....</i>	<i>25</i>
5.3.2 <i>The British Cohort Study.....</i>	<i>26</i>
5.3.3 <i>Youth Cohort Study</i>	<i>28</i>
5.3.4 <i>The Millennium Cohort Study.....</i>	<i>29</i>
5.4 INTERNATIONAL HISTORICAL LONGITUDINAL STUDIES	30
5.4.1 <i>The Historical Sample of the Population of the Netherlands.....</i>	<i>30</i>
5.4.2 <i>The linking of historic census samples for the USA.....</i>	<i>32</i>
5.4.3 <i>Swedish demographic databases.....</i>	<i>34</i>
5.4.4 <i>Programme de Recherche en Démographie Historique.....</i>	<i>35</i>
6 A POTENTIAL METHODOLOGY FOR THE VPS.....	36
6.1 SELECTING THE ORIGINAL SAMPLE MEMBERS	36
6.2 WHO REMAINS IN OBSERVATION?	38
6.3 LINKAGE.....	41
6.4 REFRESHMENT	44
6.5 PROJECT ORGANISATION	44
7 RELATED PROJECTS	46
7.1 THE HISTORICAL POPULATION REPORTS COLLECTION	46
7.2 GREAT BRITAIN HISTORICAL DATABASE.....	47
7.3 HISTORIC GEOGRAPHICAL INFORMATION SYSTEMS.....	47

8	POTENTIAL PARTNERS (IN ADDITION TO ESRC)	49
8.1	KEY PARTNERS	49
8.1.1	<i>The National Archives</i>	49
8.1.2	<i>The Office for National Statistics</i>	50
8.1.3	<i>General Registry Office, Scotland</i>	50
8.2	ASSOCIATED PARTNERS	50
8.2.1	<i>FreeDMB</i>	50
8.2.2	<i>QinetiQ</i>	50
8.3	VOLUNTEER GROUPS	50
8.4	POTENTIAL AHRB INVOLVEMENT	50
8.5	INTERNATIONAL COLLABORATIONS	51
8.5.1	<i>Minnesota Population Center</i>	51
8.5.2	<i>Historical Sample of the Population of the Netherlands</i>	51
8.6	E-SCIENCE	51
9	RECOMMENDATIONS	52
10	BIBLIOGRAPHY	54
10.1	PUBLISHED WORKS	54
10.2	UNPUBLISHED REPORTS	58
10.3	WEB SITES CONSULTED	58
10.4	DATA COLLECTIONS	60
11	INDIVIDUALS AND GROUPS CONSULTED	62

1 Executive Summary

This scoping study was commissioned by the ESRC to investigate the possibilities of creating a Victorian Panel Survey (VPS). In outline the VPS would create a significant new research resource by linking historic census and civil registration data for the period 1851 to 1901 (section 2).

By creating a unique longitudinal dataset, statistically representative of Great Britain, the VPS would open up several new research opportunities. These would cover a number of broad areas and themes in historical demography, migration studies, occupational and employment analyses, household and family structures, and the investigation of patterns in literacy and language. It would provide significant advances over existing cross sectional analyses, and create new opportunities for both historical and contemporary social scientific research, as well as for important international comparative research (section 3).

In addition to the ESRC, key stakeholders would include The National Archives, the Office for National Statistics and the General Registry Office for Scotland. The Arts and Humanities Research Board may also joint the project through the creation of enriched data sub-sets. The VPS would also drawn on volunteer researchers organised through the Federation of Family History Societies and the British Association for Local History (section 8).

A number of present day longitudinal studies are evaluated, together with relevant historical international longitudinal studies in order to assess the methodologies applied and their applicability for the proposed VPS (section 5).

The available data sources for creating a VPS are assessed (section 4) and methodological problems are discussed (section 6). Particular problems relating to the proposed VPS are identified. These include the definition of the starting population, the assignment of observation rules, record linkage strategies, and sample refreshment. In addition, possibilities regarding the organisation of the project are explored.

The potential for integrating the proposed VPS with other data resources in order to provide contextual research information is discussed (section 7) and the possibility of applying e-Science techniques is noted.

A two-year pilot project is recommended in order to further explore and test methodological, sampling and organisational issues.

2 Background

This document was commissioned by the ESRC in order to investigate the potential of creating what has been termed a *Victorian Panel Survey* (hereafter VPS). The idea of generating a VPS arose from an initiative taken by The National Archives (TNA, formerly Public Record Office) to enter into collaborative agreements with appropriate HE/FE stakeholders in order to generate new IT resources to the mutual benefit of both parties. In the discussions between TNA and the ESRC which followed, the idea of jointly creating a VPS was first raised. An outline plan was subsequently presented to the ESRC's RRB who agreed to fund this scoping study.

2.1 Rationale

The ESRC has an established track record in funding the creation of major data resources for the academic social science research community, either wholly or in partnership with other bodies. Of these data creation schemes, the generation of the British Household Panel Survey (BHPS) stands out as a major success. It is one of the most heavily used datasets in quantitative social science and has underpinned high quality research analyses and publications.¹ In part, the success of the BHPS results from the fact that it is a multifaceted data source, which can be applied to multiple research agendas across a broad range of interests within the social sciences.

No comparable resource exists within the field of economic and social history, although the need for historical longitudinal data has long been recognised. The importance of longitudinal data for social scientific research has recently been recognised by the ESRC through the funding and creation of the UK Longitudinal Studies Centre (ULSC)² and ESDS Longitudinal.³ The online database *Keeping Track*, jointly maintained by the Office for National Statistics (ONS) and ULSC, which lists and documents all 'known' longitudinal surveys currently contains 320 entries.⁴ Approximately a third of this total relate to the UK, yet of these only one is historical.⁵ Creating a nationally representative, robust and statistically significant historical longitudinal dataset for the nineteenth century would, it is believed, both stimulate and generate a substantial research output, in the same way that the BHPS and other modern longitudinal data resources have done. Such a resource would not only have relevance for social and economic historians, but would also allow social scientists working on present day issues to undertake analyses in a longer-term comparative framework.

2.2 The VPS in outline

A description of how a VPS might be created, the associated methodology and related problems are presented in a later section of this document (section 6). However, it is useful at this early stage to give a brief outline of what is currently envisaged.

¹ The BHPS is discussed in greater detail in section 5.1 of this report.

² <http://www.iser.essex.ac.uk/ulsc>

³ <http://www.esds.ac.uk/longitudinal>

⁴ <http://www.iser.essex.ac.uk/ulsc/projects/ldr4ss/index.php>

⁵ Pooley, C.G. and Turnbull, J. *Longitudinal Study of Residential Histories, 1750-1994* [computer file]. Colchester, Essex: UK Data Archive [distributor], 30 July 1997. SN: 3571. Historical in this context refers to any study earlier than 1945.

Great advances have been made in recent years in terms of the creation of computerised and searchable indices to major collections of historical sources. In relation to the proposed VPS, of particular importance are the national databases that exist for the censuses of 1881 and 1901. Following on the success of creating the 1901 internet-based resource, TNA is already committed to a programme of computerising the remaining nineteenth-century censuses for England and Wales. Likewise for Scotland, where the census has been administered separately from that of England and Wales since 1861, plans are well advanced to digitise and index all remaining unindexed censuses. In addition to this massive programme of census computerisation and indexing, the ONS has announced its intention to produce a searchable database of the 'historic' civil registers of births, marriages and deaths.⁶ Although the precise timescale for the completion of these indexing projects is not available at present, it is expected that they will be generated over the course of the next five years. In the case of Scotland, for which all the civil registration material has already been computerised, as has the 1891 census, it is reported that all the remaining censuses will have been completed by mid-2004.⁷

Although these resources are being created primarily with the huge genealogical and family history communities in mind, they potentially offer significant opportunities to the academic research community as well, particularly in facilitating the potential for tracing individuals over the course of the second half of the nineteenth century. Building on this previously unrealised potential, it is proposed to create a VPS. This would take as its base the individuals and households recorded in the existing ESRC-funded computerised national two per cent sample of the 1851 British census, created by Professor Michael Anderson of the University of Edinburgh.⁸ Using the available and nationally representative sample for 1851 as an initial 'wave' the individuals and households, recorded at this starting point would be traced through the subsequent registration and census information for the fifty-year period to 1901. The result would be a linked database with each census year between 1851 and 1901 in essence acting as a surrogate 'wave'.

As mentioned already, the proposed methodology will be detailed in a later section, however, it is clear that the success (or otherwise) of the proposed VPS project will to a large extent depend on the ability to trace and link individuals over time. Even with the national indices for the censuses and registration materials, this will be a daunting task in terms of the effort involved. As a result, it is proposed that the project will call upon and work with the large and active genealogical, family historian and local history communities. These will serve as volunteer researchers tracing the original household members and 'incoming' household members between the censuses, with a centralised research unit

⁶ See ONS, *Civil registration*. It is not known at this stage what will be included within the 'historical' category, and this will most probably vary depending on the nature of the event, but certainly all records relevant to the VPS would be 'historical'.

⁷ Spencer, 'Scotlands people web site', 55.

⁸ ESRC award RH00232032. The data resulting from the project are available as Anderson, M. et al. *National Sample from the 1851 Census of Great Britain* [computer file]. Colchester, Essex: UK Data Archive [distributor], 1979. SN: 1316. The original data deposited with the UK data Archive have subsequently been completely checked, cleaned, recoded and reformatted at the Department of History, University of Essex to make them directly comparable with the available data for the 1881 census. See Schürer, K. and Woollard, M. *1881 Census for England and Wales, the Channel Islands and the Isle of Man (Enriched Version)* [computer file]. Genealogical Society of Utah, Federation of Family History Societies [original data producers]. Colchester, Essex: UK Data Archive [distributor], 2000. SN:4177 and Schürer, K. and Woollard, M. *1881 Census for Scotland (Enriched Version)* [computer file]. Genealogical Society of Utah, Federation of Family History Societies [original data producers]. Colchester, Essex: UK Data Archive [distributor], 2000. SN:4178.

serving to co-ordinate the volunteer effort and acting to ensure standardisation and quality control. In this regard, the active collaboration and partnership with TNA is seen as fundamental. Not only is TNA the major historical archive in England and Wales, it also runs in partnership with the ONS, the Family Records Centre (FRC) acknowledged as the main source of information for genealogists and family historians researching the nineteenth century. Using TNA's existing and extensive network it is planned to harness the input to the project of a significant body of volunteer researchers.

3 Research Issues

3.1 Potential use of a VPS

There is little point in the ESRC sponsoring the creation of any new resource in the social sciences unless it can be justified in terms of the quality of the research than will be generated from it, either direct or indirectly. Trying to estimate the likely use of a new resource prior to its successful creation and dissemination will always be prone to a certain amount of speculation. This is especially true in the case of the VPS where no comparable historical resource currently exists.

In trying to establish the need for a resource such as the VPS, the degree to which it may be used and the research issues it may be utilised to address, it is instructive, first, to look at the established track record of contemporary data resources that might be seen as being complimentary.

3.2 The evidence of modern longitudinal studies

In relation to the BHPS, which started in 1991 and has conducted subsequent follow-up waves annually thereafter, it is stated that a longitudinal panel study has a number of advantages over other cross-sectional surveys, as follows:

- they allow analysis of how individuals and households experience change in their socio-economic environment and how they respond to such changes;
- they allow an analysis of how conditions, life events, behaviour and values are linked with each other dynamically over time;
- they allow analysts to control for unobserved heterogeneity in cross-sectional models through difference analysis;
- because all household members are interviewed, the effects of the interaction of changes at the individual level can be analysed for the whole household or for other individuals;
- because sample members are followed as they leave their original household. Panel data provide unique information on the processes of household formation and dissolution.⁹

The BHPS is certainly heavily used by academic social scientists. Over the past decade it has consistently been one of the most requested datasets disseminated by the UK Data Archive (UKDA), and the most heavily used dataset generated as a result of ESRC funding. In the past three years it has been ranked, respectively, as the 4th, 2nd and 4th most popular dataset disseminated by the UKDA.¹⁰ Details of its usage statistics for the period 1999 to 2003 is given in Table 1, together with other major longitudinal studies disseminated by the UKDA.

⁹ This section of the report draws upon Taylor *et al*, *BHPS user manual*, especially section II. However, it is important to remember that the BHPS updates annually, whereas this is clearly not possible for the VPS, thus the VPS would never have the richness of dynamic data captured by the BHPS.

¹⁰ Indeed in the reporting year 2001-02 there were five longitudinal data series recorded in the top twenty datasets disseminated by the UKDA. In addition to the BHPS, BCS70 was the 7th most requested data series, YCS 13th, NCDS/BCS 14th and NCDS 15th.

Table 1. Usage of major longitudinal data series

	2000-01	2001-02	2002-03
	n. of users	n. of users	n. of users
BHPS	155	375	429
NCDS	41	74	89
BCS70	32	92	195
NCDS/BCS70	2	89	126
YCS	13	41	53
MCS	N/a	N/a	18

Note: Year runs from Sep 1 to Aug 30. There may be some double counting where users have both ordered data (portable, access) and downloaded data (via download) in the same year.

This use has translated itself into a large number of research publications and outputs. To date a total of some 690 publications are known to have resulted from, or are linked to, analyses of the BHPS.¹¹ Alongside the BHPS, it is possible to examine the past usage of the Longitudinal Study (LS), a survey initiated and managed by ONS, but for which ESRC support has also been made. The LS is in fact more akin in some respects to the proposed VPS, given it fundamentally links census data from 1971, 1981 and 1991 through intervening registration data.¹² The web site of the Centre for Longitudinal Studies, who facilitate academic access to the LS and monitor its use, currently list a total of 576 publications arising from work undertaken using the LS. These are categorised into the subject groups or themes given in Table 2. This shows the importance of health studies in using the LS, mainly resulting from the fact that the LS, as would the VPS, records the cause of death of the subjects that it traces, in the event that death has occurred.

Table 2. Numbers of publications arising from research using the Longitudinal Study

Cancer studies	82
Geographic differences in health and mortality	28
Inequalities in women's health and mortality	62
Socio-economic differences in health and mortality	123
Ethnicity	37
Gerontological studies	19
Migration	69
Demographic studies	65
Social and economic change	37
General and methodological	54
Total	576

3.3 Potential applications for the VPS

The manuscript nineteenth-century census returns, usually referred to as the census enumerators' books (CEBs), have previously been heavily used in historical research. Indeed, they could be said to form the backbone of many strands of investigation within social and economic history for the nineteenth

¹¹ <http://www.iser.essex.ac.uk/bhps/doc/vola/app5.html>

¹² Details of the methodology underlying the LS are given in section 5.2 of this report.

century. An annotated bibliography of research based substantially on the CEBs published in 1989 lists some 423 publications.¹³ These can be subdivided into the thematic categories listed in Table 3. In the fifteen years that have passed since this list was compiled the number of published studies has clearly increased, since the popularity of the CEBs as a key source shows no sign of diminishing. The current figure of published outcomes from the CEBs is now possibly double that reported in 1989, at around 850, but there is no reason to suggest that the overall balance in terms of subject matter is fundamentally different from that recorded in Table 3. It is also important to note that these figures take no account of unpublished masters and doctoral research, for which the CEBs have been extensively used within the field of nineteenth-century social and economic history.

Table 3. Topics investigated in the CEB publications listed in Mills and Pearce (1989)

Occupations	257
Migration	218
Demographic	168
Household	160
Methodology	122
Segregation	100
Social structure	69
Family	66

As with research on modern day populations, there are many potential research questions that historical longitudinal can address which cross-sectional data, such as recorded in the CEBs, simple cannot. Indeed, it is because of the potential offered by longitudinal data that a number of researchers have already constructed longitudinal databases based on the CEBs. Research using linked CEBs has already made important contributions and offered valuable insights into a number of areas of social and economic history. These have included work on historical demography,¹⁴ migration,¹⁵ occupational structures,¹⁶ household composition,¹⁷ as well as evaluations of the accuracy of the underlying source materials.¹⁸ However, the existing body of nineteenth-century longitudinal data using the CEBs, and the studies arising from them all suffer from two common and fundamental methodological problems.

First, out of practical necessity they are all essentially 'place orientated', based on local populations. Previously, the basic method of constructing a longitudinal database from the CEBs has been to take the records of individuals and households for one particular place (usually a parish) and to link them incrementally to the CEBs for subsequent censuses for the same place. As a result a large number of individuals are lost from observation (potentially up to forty per cent, and often more in large urban areas) as they move away from the place

¹³ Mills and Pearce, *People and places*, 6. However the number of separate publications is a little under 400 since some publications are listed more than once due to the nature of their geographical coverage.

¹⁴ See, for example, Hinde, 'Population of a Wiltshire village'; Garrett, 'Trials of labour'; Reay, 'Before the transition'.

¹⁵ See, for example, Dennis, 'Distance'; Dennis, 'Intercensal mobility'; Pooley, 'Residential mobility'; Pooley and Doherty, 'The longitudinal study of migration'; Schürer, 'The role of the family'; White, 'Family migration'; Wojciechowska, 'Brenchley'.

¹⁶ See, for example, Crompton, 'An exploration'; Hallas, 'The social and economic impact';

¹⁷ See for example, Hancock, 'In service'; Nenadic, 'Studying the middle class'; Reay, 'Kinship and the neighbourhood'.

¹⁸ See, for example, Perkyns, 'Age checkability'; Perkyns, 'Birthplace accuracy'.

being studied between censuses. Thus the resulting database and research results derived from it are essentially those for a stable or non-migratory subset of the population.¹⁹

Second, since access to the civil registration material has not been available in a useable and effective form, previous studies constructing longitudinal databases have tended to link census to census. This not only reduces the number of links possible,²⁰ but also means that intervening events are outside of observation. The inability to link the civil registration data to the CEBs also reduces the range of information that can be associated with an individual's 'life history' record.

The creation of a statistically representative national longitudinal sample, such as the VPS, in which the linking of individuals was undertaken at a national rather than local level, using both census *and* civil registration indices, would provide a workable solution to both of these existing methodological problems. At the same time, a standardised national sample would facilitate comparative research, at a variety of levels:

- geographical comparison between regions;
- comparison over time, contrasting the experience of the VPS with present data longitudinal studies;
- international comparisons, contrasting the British experience with that of other countries.

It is clearly not possible to predict all the areas of research that the VPS, if created, might be used to address, but the following list, subdivided into broad thematic categories, provides an initial indication of potential areas of investigation, none of which can be addressed satisfactorily with existing cross-sectional resources.

Demography

- The analysis of demographic events (fertility, nuptiality and mortality) within the dynamic household and familial contexts in which they occur.
- The construction of cohort demographic rates in order to re-evaluate existing knowledge of the demographic transition, most demographic work for the nineteenth century previously being based on period rates.
- The construction of cohort life tables by occupation groups and social class.
- Contrasting the demographic experience of migratory and non-migratory populations.
- The analysis of demographic experience by religious denomination.
- The analysis of cause of death within the context of dynamic household and family structures.

¹⁹ A similar problem, of course, exists in the case of family reconstitutions conducted for earlier periods using parish registers in order to calculate historical demographic rates. See Wrigley *et al*, *English population history*.

²⁰ In particular, single women may be lost from observation following marriage due to them changing surname, and children who are both born and die between censuses are lost. Researchers have used local parish registers to overcome these problems, but due to the incompleteness of ecclesiastical registration for most places in the second half of the nineteenth century, this offers only a partial solution at best.

- The investigation of generational effects on nuptiality, fertility and mortality (the extent to which a parent's demographic experience, influences that of their children).
- The analysis of demographic experience by literacy/illiteracy.

Migration

- The reconstruction of macro-migration experiences over the life-course in the context of the timing of demographic events.
- The analysis of regional migration flows and changing regional economic structures, in a socio-economic context.

Emigration

- The construction of emigration rates and estimates.
- The analysis of the demographic, social and economic profile of emigrants.

Household and family

- The investigation of the processes of household evolution and dissolution within regional, social and economic contexts.
- The examination of non-family members (servants, lodgers, boarders) and their movement into and out of 'normal' family groups.
- The analysis of household structures by religious denomination.
- The examination of individuals moving to and from familial patterns of residence to institutional living arrangements (e.g. workhouses).

Employment

- The analysis of occupational mobility.
- The analysis of social mobility.
- The examination of changing household-based work histories.
- The validity and consistency of occupation recording.

Literacy and language

- The examination of literacy in the context of generational, regional, social and economic indicators.
- The investigation of Gaelic and Welsh speaking populations in a longitudinal context.

General

- The examination of regional change with respect to (for instance) occupations, household structure, birthplace/ethnicity.

- The analysis of the experience of change over time for particular sub-groups of the population. For instance: particular occupational groups, ethnic (birthplace) groups, those at a particular stage of the life cycle, those in particular household positions. Thus it would be possible to focus (for example) on married women age 20-29 and examine their occupations and family circumstances over a 50-year period.
- The development of vignettes, or life histories, of particular groups within the population as illustrations of broader trends.
- The investigation at a national and regional scale of a whole range of questions that are usually only investigated at a local scale. For instance, examination of changes in detailed household structure with regard to the position of servants and lodgers.

However, despite the many new and important research issues the VPS would help to address, as with any major data resource, historical or contemporary, it will not be without limitations. Depending on the sampling frame adopted, it will not be possible to use the VPS in certain ways, as is the case with the BHPS and LS, for example. This point is discussed further in the Methodology section of this report. Also, as with all large-scale quantitative data sources, it will always benefit from being analysed alongside and in combination with complimentary qualitative sources. Equally, being essentially a product of the nineteenth-century census and registration systems it will fail to capture information on events and processes that these sources do not witness (for example, inter-census migration or occupation change not otherwise recorded in any intervening civil registration events). Lastly, the extent to which analyses can be undertaken on small sub-groups of the population, for example some occupational groups or minority language speakers, will depend on the overall size of the VPS. This last point is also further discussed in the Methodology section of this report.

4 Sources

4.1 Historical census data

The first census in Great Britain was taken in 1801.²¹ This census and its three successors were organised by John Rickman and undertaken by overseers of the poor, and members of the clergy in England and by schoolmasters in Scotland. These early censuses not only enumerated the current population but also attempted through the use of abstracting material from the parish registers to show the changes in the population of the country throughout the eighteenth-century.

The administrative task of taking the census was transferred to the General Register Office (GRO, see section 4.2 below) in time for the 1841 census. For England and Wales the GRO was to continue to take the census until 1970 when it merged with the Government Social Survey Department to form the Office of Population Censuses and Surveys (OPCS). This body was subsequently merged into the Office for National Statistics (ONS) who oversaw the taking of the census of 2001.

Until 1861 the Census in Scotland was the responsibility of the GRO when it then became a function of the newly created Registrar-General for Scotland (GRO(S)). The GRO(S) was created in 1854 after the passing of an Act 'to provide for the better registration of births, deaths and marriages in Scotland'. GRO(S) has been responsible for taking the census in Scotland since 1861, though provision for these censuses was, after 1920, made through a unified Act of Parliament.

The first four national censuses (1801-1831) primarily recorded information in aggregate form only. The process of census enumeration was fundamentally changed for the census of 1841 when householders' schedules were introduced. As part of this process the census enumerators provided each householder with a *pro forma* schedule to be completed for census night (always a Sunday night, 1841-1901), following the instructions printed on the schedule. The schedules were then collected by the enumerator on the following day and copied into printed census enumeration books (CEBs). Following completion, these CEBs were sent to the Census Office in London for checking, abstraction and analyses. The original householders' schedules were later destroyed and it is the CEBs that survive as public records for historical enquiry, and that will form the backbone of the proposed VPS.

The 1841 census was the first to record the names of individuals in a routine fashion. However, the 1841 census was limited in the amount of information it collected on individuals, while from 1851 onwards the census collected a much wider range of demographic, social and economic materials. Thus, the census of 1851 forms a logical starting point for the proposed VPS. The types of information collected by the censuses of 1851 to 1901 are summarised in Table 4 below. It is important to note that in most respects the English and Welsh, and Scottish censuses collected similar information, though at times the Scottish census collected additional information on language and housing conditions.

²¹ For background information on the historic censuses see Higgs, *Making sense*; Higgs, *Clearer sense*; Lawton, *Census and social structure*; and Mills and Schürer, *Local communities*.

Table 4. Information recorded in the CEBs, 1851 to 1901

Description of column	Variations
Number of schedule	In all years
Address information	In all years
Houses, inhabited, uninhabited and building	In 1901 inhabited; in occupation, not in occupation and building
Number of rooms occupied if less than five	In England and Wales only, for 1891 and 1901
Number of rooms with one or more windows	In Scotland only, for 1861, 1871, 1881, 1891 and 1901
Name and surname	In all years
Relation to head of family	In all years
Condition (as to marriage)	In all years
Age	In all years, but in 1881-1901 column was changed to read 'age last birthday'
Rank, profession or occupation	In 1891 and 1901 'rank' was dropped
Employer, employed, neither employer nor employed (three columns)	In 1891 only
Employer, worker, or own account	In 1901 only
If working at home	In 1901 only
Education: if receiving regular instruction at home	In Scotland only, for 1881, 1891 and 1901
Where born	For those born in the British Isles, usually in the form of parish and county of birth. For those born overseas, normally country only recorded
Whether deaf-and-dumb, blind, imbecile or idiot, lunatic	Limited to blind and deaf-and-dumb in 1851 and 1861. In 1901 'feeble-minded' instead of 'idiot'
Language spoken	In Wales a question was asked in 1891 and 1901 if the respondent spoke English, Welsh, or both, the question being limited to. Children under three excluded in 1901 A similar question as regard to Gaelic was asked in Scotland in 1881, 1891 and 1901.

It is important to add a note on the situation regarding census taking in Ireland. The first properly administered census was taken in Ireland in 1821 and subsequent censuses were administered by a group of Census Commissioners. The censuses of 1821 and 1831 were little more than headcounts, but subsequent

censuses collected a wide range of material leading to their description as social surveys rather than simple population counts. The 1841 census was led by Captain Thomas Larcom, but the 1851 census was taken by the newly created Registrar-General for Marriages, which in 1857 formally took responsibility for the census until 1911. No census was taken in Ireland in 1921, but both the Republic and Northern Ireland took a census in 1926.²² However, most of the census returns for the nineteenth century were destroyed in a fire in 1922, and only those for 1901 and 1911 survive in entirety. Thus, regrettably, due to the lack of source materials, it is not possible to include Ireland in the VPS.

A lot of nineteenth-century CEB data have already been transcribed and computerised. In addition to the 1851 national sample created by Michael Anderson, a one hundred per cent version of the 1881 censuses of England and Wales and Scotland exists in database form. This has been coded and enhanced at the University of Essex and has been deposited with the UKDA for academic use.²³ More recently, the Essex team has re-coded and formatted the original 1851 sample in order to make a new version which can be used for comparative research alongside the 1881 database.

In 2002 a fully searchable internet-based database of the 1901 census for England and Wales was made available by QinetiQ in association with TNA. Similar databases exist for Scotland for 1891 and 1901, created as a result of a partnership between the GRO(S) and Scotland On Line. It is planned to complete the Scottish censuses of 1861 and 1871 by Easter 2004. Equally, TNA has plans to computerise the censuses of 1861, 1871 and 1891, but a definite timetable has yet to be announced. Lastly, a charitable organisation FreeCEN are recruiting volunteers to transcribe and computerise nineteenth-century CEBs and make them freely available on the internet. Efforts to date have concentrated on 1891.

4.2 Historical registration data

The General Register Office (GRO) was created in 1837 as a result of the Births and Deaths Registration Act, 1836. Its initial responsibility was to oversee the civil registration of births, deaths and marriages in England and Wales, and to publish statistics based on these registrations. From 1837 to 1871 the GRO was a subordinate department of the Home Office, when it passed to the Local Government Board. In 1919 responsibility for the GRO transferred to the Ministry of Health, and it is now under the auspices of the ONS. The system of civil registration which started in England and Wales in 1837 did not commence in Scotland until 1855.²⁴

The published Annual Reports and Decennial Supplements of the various Registrars-General provide aggregated statistics on the recorded numbers of births, marriages and deaths, as well as cause of death, for a variety (and inconsistent number) of administrative units. These have proved to be a prime source for research on the demographic and medical history of the nineteenth

²² See Crawford, 'Counting the people' and Royle, 'Irish manuscript census records'.

²³ Schürer, K. and Woollard, M. *1881 Census for England and Wales, the Channel Islands and the Isle of Man (Enriched Version)* [computer file]. Genealogical Society of Utah, Federation of Family History Societies [original data producers]. Colchester, Essex: UK Data Archive [distributor], 2000. SN:4177 and Schürer, K. and Woollard, M. *1881 Census for Scotland (Enriched Version)* [computer file]. Genealogical Society of Utah, Federation of Family History Societies [original data producers]. Colchester, Essex: UK Data Archive [distributor], 2000. SN:4178.

²⁴ For information on the GRO, see Nissel, *People count*.

century.²⁵ However, despite the richness of information contained within the certificates of births, marriages and deaths, these have been rarely used by academic researchers. This is primarily a result of the way in which access to the individual certificates is currently organised. In England and Wales the names of the child (for births), marriage partners and those having died are recorded, along with the place where the event was registered, and indexed chronologically, by year and quarter. These indices are openly available for searching at the Family Records Centre in London but only provide the reference details to the certificate which needs to be consulted in order to ascertain the full details of the registration event in question. Using these reference details, a copy of the given certificate then has to be ordered, which takes four working days to process, and costs £7 per individual certificate. This effectively rules out any historical research which may require information from hundreds or thousands of certificates to be transcribed into database form.

Yet the current system for accessing the information recorded on the certificates generated by the civil registration process is set to change. Plans recently released in a public consultation document on proposed changes to the legislation relating to the Civil Registration Service in England and Wales include proposals which will radically transform access to the information contained within the civil registers for the nineteenth century.²⁶ This report firstly recommends that a distinction be drawn between historic records (generally those over 100 years old) and modern records. Having made this distinction, it is then proposed that the details from the historic certificates are computerised and held in a central database. In order to improve access to these historic records, the report proposes to make searchable access to the database of historic records 'available over the Internet, probably free of charge'.²⁷ Discussions with ONS in the course of preparing this report suggest that this proposal is almost certain to be accepted, however, the precise timing is uncertain.

In Scotland the GRO(S), in partnership with Scotland On Line, have already computerised the civil registers and created an internet-based searchable database. This is a great improvement on the current situation for England and Wales, but researchers still have to pay to access the database, with costs dependant on the time spent on the system, rather than the number of entries searched.

The liberation of the civil registration information will be a massive bonus not only to genealogists, but also for academic researchers. It paves the way, for the first time, for large-scale research to be undertaken on the information contained within the individual certificates rather than on the published aggregate tables. Since most historians and social scientists have not previously used the certificates generated by the civil registration process for research, it is worth noting the range of information that they contain.²⁸

Information on marriage certificates

- date of marriage
- place of marriage

²⁵ See, for example, Anderson, 'Marriage patterns'; Baines, *Migration*; Friedlander and Roshier, 'Study of internal migration'; Lawton, 'Population changes'; Tetielbaum, *British fertility decline*; Woods, 'Approaches'; Woods, *The demography*.

²⁶ ONS, *Civil registration*.

²⁷ ONS, *Civil registration*, section 6.4.13.

²⁸ For background information on the registers see Annal, *Using birth, marriage and death records*, and for more detailed information on the recording of the material contained within the registers see Dixon, *Marriages* and Dixon, *Birth and death*.

- whether by banns, licence or certificate
- religious order in which the marriage was solemnised
- name of the bride
- name of the groom
- age of the bride
- age of the groom
- current address of the bride
- current address of the groom
- marital status of the bride
- marital status of the groom
- occupation of the bride
- occupation of the groom
- name of the father of the bride
- name of the father of the groom
- occupation of the father of the bride
- occupation of the father of the groom
- names of two or more witnesses
- name of the person(s) performing the ceremony

Information on death certificates

- date of death
- place of death
- name
- age at death
- cause of death
- occupation (or name and occupation of her husband, if the deceased was a married or widowed woman)
- name, address and relationship to the deceased any) of the person who registered the death

Information on birth certificates

- date of birth
- place of birth
- child's forenames
- sex
- father's name
- father's occupation
- mother's name and maiden name
- name, address and relationship to the child of the person who registered the birth

Lastly, it is also worth noting that some county family history societies have already transcribed the civil registration indices at a local level. Of particular note, however, is the FreeBMD project. Calling on an army of volunteers from within the genealogical community this project has already transcribed 51.3 million entries for the period 1851 to 1901 from the indices of the civil registers for England and Wales. As shown in Table 5, below, this figure represents some 58 per cent of the total of 88.4 million entries to be transcribed.²⁹ The coverage is currently rather uneven with the 1881-1900 period being rather better covered than that of 1851-1865, but importantly marriages, which are the most significant for developing linking software due to the surname change of women, are generally very well covered. FreeBMD estimate that the task of transcribing the civil

²⁹ Marriages are counted twice since these result in two entries in the marriage index, for both bride and groom.

register indices will be completed during the course of 2004. All the transcribed entries are freely searchable on the internet.

Table 5. Proportion of events transcribed by FreeBMD

	Births	Marriages	Deaths
1851-55	33.5	49.3	41.0
1856-60	35.6	24.4	48.0
1861-65	53.2	1.2	4.7
1866-70	9.6	100.0	41.9
1871-75	85.6	90.6	52.5
1876-80	35.9	85.4	40.8
1881-85	25.4	95.7	64.9
1886-90	21.3	99.7	36.6
1891-95	85.1	100.0	99.5
1896-00	97.4	96.4	83.2
1851-1901	50.4	79.2	53.9

4.3 The 1851 census sample

The national two per cent sample of the 1851 census of Great Britain created by Michael Anderson is a stratified systematic cluster sample.³⁰ In the case of the 1851 census (as with all censuses from 1841 to 1991) the smallest unit of administration was the enumeration district (ED). This was the area covered by a single census enumerator and was limited in terms of physical size to an area for which a single person could collect all the distributed census householders' schedules in a single day. In drawing up the boundaries of EDs instructions were given that these should be consistent, as far as practically possible, with other existing administrative boundaries. Thus in rural areas EDs were usually coterminous with civil parishes, while large urban areas were sub-divided into multiple EDs. Liverpool, for example was split into some 300 districts.³¹

The individuals and households captured in the 1851 national sample are clustered into EDs since, at the specific request of the then SSRC who funded the creation of the sample, the ED, rather than the individual or household, was designated as the smallest sampling unit. In total, 980 EDs, or 'data clusters', were captured. However, it was also decided to stratify these clusters according to different typologies. In order to maximise the extent to which the sample was representative of different kinds of places, all settlements listed in the published census reports were first stratified or divided into five main groups:

- towns;
- small non-urban settlements;
- large non-urban settlements;
- other places;
- institutions.

³⁰ This part of the report is based on the section on sampling in Anderson, *National sample*.

³¹ R. Lawton, 'Census data for urban areas', 127.

The first of these, 'towns', consists of all the settlements located within the boundaries of Municipal Boroughs and of other places identified as 'towns' by the census authorities. Settlements within Parliamentary Boroughs were also included in this 'urban' category. The second category, 'small non-urban settlements', consists of all settlements listed in the published census reports with a population in 1851 of 2,000 or less. All remaining complete settlements were classified as 'large non-urban', except for a few areas within settlements which were outside town or borough boundaries; these made up the residual 'other places' category. The last category, 'Institutions' includes all those establishments listed as such in the published census reports. These include such establishments as workhouses, barracks, ships, prisons and were recorded in special enumeration books, thus the identification of these from a sampling point of view is relatively straightforward.³²

Having defined the stratifications, the sample was then drawn, with two major exceptions, by selecting every fiftieth enumeration book within each defined category of place. The first exception was that, in England and Wales, the small non-urban settlements were sampled taking the whole of every fiftieth appropriate place as listed in the published census Reports, thus capturing whole villages.³³ The second exception was made in the case of institutions which were sampled by treating all the institutional populations as if they were one continuous list and then systematically selecting twenty individuals from each successive one thousand names. Where institutions included families, special arrangements were made; in these cases the sample was drawn so as to include all members of families the first member of which appeared before the twentieth individual, and to exclude members of families where only the later members appeared within a block of twenty names. The key characteristics of the 1851 sample in terms of the populations covered by stratification group are given in Table 5.

Table 5. Size of the 1851 sample by stratification type

Stratification type	Population	n. of households
'Communities' under 2,000 population	124,819	25,659
'Towns' and municipal Boroughs	157,080	34,121
Parliamentary Boroughs	27,287	5,583
'Large non-urban communities'	66,837	13,550
Residual 'non-urban' areas	9,779	1,933
Unallocable 'urban' areas	7,836	1,624
Institutions	4,763	*323
Total	398,401	**82,470

* This figure is the total number of institutions in the sample

** This figure excludes the 323 institutions

³² The official census definition of what constituted an institution was problematic. In 1851 institutions with over 200 persons should have been automatically enumerated in the separate institutional books. However, size was not the only criterion used. Certain 'standard' types of institutions (e.g. ships, workhouses, prisons, barracks) should have been entered in the special institutional books regardless of the number of inmates and staff they housed, but this was not always the case. Likewise, there was confusion over where other 'quasi-institutional' establishments such as schools, boarding houses and factories should be recorded, especially when these were smaller than the official size. See Higgs, *Making sense*, 37-9; Higgs, *Clearer sense*, 35-7; Taylor, 'Liverpool's institutional'.

³³ Although, most smaller villages or rural communities were, in 1851, enumerated within a single ED anyway, so this is not a major change.

5 Methodology utilised in existing longitudinal studies

In specifying a methodology for the proposed VPS, it is worthwhile, to consider the underlying methodologies that have been adopted in other longitudinal studies. This is instructive for two main reasons: first, in learning from the experience of other existing studies; and second, to try to ensure, wherever possible and appropriate, that the methodology adopted for the VPS allows for comparative analyses to be undertaken, and allows research outcomes from the VPS to be compared to similar findings from other longitudinal data studies. Consequently, the following sections review, in outline, the methodological approaches taken by existing longitudinal studies.

5.1 The British Household Panel Survey

The British Household Panel Survey (BHPS) is a longitudinal survey of private households in Great Britain, which started in 1991 and has grown with successive annual waves since that date. The BHPS is being carried out by the Institute for Social and Economic Research (incorporating the ESRC Research Centre on Micro-Social Change), at the University of Essex and is funded by the ESRC with support from a number of government departments and agencies.³⁴ The main objective of the survey is to further understanding of social and economic change at the individual and household level in Britain, and to identify, model and forecast such changes, their causes and consequences in relation to a range of socio-economic variables.

The initial selection of households for inclusion in the panel survey was made using a two-stage stratified systematic method as a balance between efficiency and cost and is approximately equivalent to the current sample design of the General Household Survey (GHS). This sample design is an approximately equal probability of selection method (espem) design. The sampling frame used for the selection of the sample units was the 'small users' Postcode Address File (PAF) for Great Britain south of the Caledonian canal (i.e. excluding Northern Ireland). This is the main sampling frame used by most large government surveys and is generally accepted as being the most practical and robust for the selection of samples of households in Great Britain. In the first stage of selection 250 postcode sectors were selected as the primary sampling units (PSUs) from an implicitly stratified listing of all sectors on the PAF using a systematic sampling method. In the second stage of selection, delivery points, which are approximately equivalent to addresses, were sampled from each selected PSU using an analogous systematic procedure.

Via this procedure, a sample of 8,217 addresses was drawn. For the initial wave interviews were attempted at all private households found at these addresses. All individuals enumerated in respondent households became part of the longitudinal sample. All these sample members are known as Original Sample Members (OSMs). Subsequent waves consist of all adults (those aged 16+) in all households containing at least one member who was resident in a household interviewed at wave one. This was regardless of whether individuals had been interviewed in wave one. Thus an attempt was made to interview individuals in responding households who had refused to participate in Wave One, or had been unable to take part. After wave one Original Sample Members (OSMs) were followed into non-private households (institutions) and into Scotland north of the

³⁴ Including the Health Education Authority, ONS and Eurostat.

Caledonian Canal, however, there was no attempt to follow members who had entered prison, or members who had become too frail or mentally ill. Children are interviewed once they reach the age of 16; and there was also a special survey of 11 to 15 year old household members from Wave Four onwards. Thus the sample remains broadly representative of the population of Britain as it changed through the course of the 1990s. The following were included as targets at each wave subsequent to the first:

- children born to OSMs;
- OSMs who move into a 'new' household with one or more new people;
- individuals who move into an 'existing' household with an OSM.

Children born to OSMs after the start of the study automatically become OSMs in their own right. New entrants, those in the second and third categories above, become eligible for interview as 'attribute' to an OSM, but do not themselves become an OSM. Interviewees who do not qualify as OSMs are only re-interviewed in subsequent years if they are still co-resident in a household with an OSM.

In order to maximise linkage between waves, and keep as many OSMs in observation as possible, if an OSM is not found at the expected address interviewers will attempt to trace them. This tracing is aided by the compilation of an extensive database of information on respondents' location. Additionally, after interview a thank-you letter, gift voucher and change-of-address card is sent to the interviewee. Prior to the next wave a summary of results and an address confirmation card is sent to all adults. These measures allow records to be continually updated, alert interviewers to potential problems and minimise the tracing work required of them.

At wave nine two additional samples from Scotland and Wales were added to the BHPS, in order to both boost the relatively small number of sampled households within these two countries and to facilitate analysis of the two countries compared to England in order to assess the impacts of following on from devolution. A consultation period in the early part of 1999 established the requirements of the Scottish and Welsh user-communities. Provision of comparable data between the different parts of Great Britain required identical questionnaires and fieldwork arrangements for the additional samples to those used for the main BHPS sample. The target sample size in each country was 1,500 households. The Scottish sample includes the population living north and west of the Caledonian Canal. All members of households recruited at the first wave of the extension sample are treated as OSMs, and standard BHPS apply thereafter. Members of these samples who move to England will be followed.

At wave eleven a substantial new sample in Northern Ireland, the Northern Ireland Household Panel Survey (NIHPS) was added, jointly funded by the ESRC and government departments in Northern Ireland. Since the start of the BHPS it was recognised that a sample was needed for Northern Ireland so that the coverage of the panel was UK-wide rather than for Great Britain only. Provision of data for Northern Ireland which are comparable with data from other parts of the UK required largely identical questionnaires and fieldwork arrangements for the NIHPS to those used for the other BHPS samples. The target sample size was 2,000 households. All members of households recruited at the first wave of the NIHPS are treated as OSMs, and standard BHPS following rules apply thereafter. Members of these samples who move to other parts of the UK will remain in sample.

5.2 The Longitudinal Study

The LS was designed as a continuous, multi-cohort study using one of four possible annual birth dates as the selection criterion.³⁵ This has provided a random, one per cent sample of the population of England and Wales clustered by date of birth. The initial sample was drawn from every person enumerated in the 1971 census, including visitors, and these have been subsequently traced to the 1981 and 1991 censuses. Between censuses new members, with the same sample dates of birth, enter the sample via birth or immigration. Members exit the sample through death or emigration. However, the additions and deductions through immigration and emigration are known to be incomplete. To this extent the LS, by its own admission, departs from a true population sample. Tracing members of the sample population from census to census and the linking in of vital events was facilitated by utilising the National Health Service Central Register (NHSCR) and the 'person' unique NHS number. This is the only commonly held means of identification. Unlike other means of identification, such as the National Insurance number, the NHS number is issued at birth registration and, therefore, allows linkage events to children as well as adults.

The LS is designed to be a sample with a continuous updating of population. The components of population change, births, deaths, immigration and emigration, are added or subtracted as necessary. These events are notified to, and recorded by, the NHSCR. The recording of births and deaths stands at nearly one hundred per cent, but the notification of immigration and emigration is substantially lower.

An exit event (emigration or death) does not mean that a member is removed from the files. The event is recorded for that member. An emigrant may later re-enter the country and become re-activated.

In summary, the LS is updated when the following events occur to sample members:

- births to sample members;
- infant mortality to babies of sampled mothers;
- death of the spouse of a LS member;
- cancer registration;
- enlistment into the armed forces;
- entries into long-stay psychiatric hospitals;
- exits from psychiatric hospitals and armed forces.

Since information is recorded on those who are co-resident with LS sample members, as recorded in the censuses, it is important to realise that the LS can also be used as a quasi-panel study, for example, to analyse changes in household composition overtime.

Initially the LS study covered England and Wales, with a parallel study running in Scotland. After 1981 the LS covered only England and Wales, the Scottish LS being abandoned due to financial restraints.

The linkage between the 1971 and 1981 censuses, and the additions and subtractions carried out between these dates, was carried out manually. After 1991, when the NHSCR was computerised, linkage, additions and subtractions were carried out automatically. Ninety-eight per cent of the 1991 sample members, who were extracted from the census of that year, were identified by NHS numbers at the NHSCR. Of these, seventy-nine per cent could be linked to the 1981 census. Fourteen per cent were new births or immigrants. Three per

³⁵ This section draws on Hattersley and Creeser, *Longitudinal study*.

cent were enumerated in 1971, but missed in 1981. Four per cent were new entrants who did not come into any of the above categories. Between the 1981 and 1991 censuses a linkage rate of ninety per cent was achieved.

5.3 Other UK cohort studies

5.3.1 The National Child Development Survey

The National Child Development Survey (NCDS) was instigated to monitor the physical, educational and social development of those people born in Britain in the week beginning 3 March 1958.³⁶ There have been six attempts to trace all the members of this birth cohort in 1965, 1969, 1974, 1981, 1985 and 1999/2000. The 1999/2000 sweep was combined with the latest wave of the 1970 British Cohort Study (see section 5.3.2).

The primary means of gathering data in these sweeps was via interviews, questionnaires and tests for the children. The information on the birth was obtained from the mother and from the medical records of the midwife. The first three sweeps gathered information from parents, head teachers, class teachers and the testing of the subjects. In 1981 the subjects themselves were interviewed. Additional information was taken from the 1971 and 1981 censuses.

In 1991 the cohort members were interviewed along with their spouses, cohabitees and children (as appropriate). In this survey the mother and child questionnaires were based on the instruments used for the United States National Longitudinal Study of Youth to allow comparisons to be made.

The original birth cohort was augmented by the inclusion of immigrants born in the same week as the target sample. These immigrants were identified from school registers during the process of tracing the original members. This was only carried out during the first three sweeps. Since 1974 no attempt has been made to include new immigrants in the survey. Table 6, below, contains the details of the collection of data for each sweep.

Table 6. NCDS: details of survey and collection methods used

Survey	Number	Source of information	Method of data collection
PMS (17,414)	17,414	Mother Medical records	Interviewed by midwife Consulted by midwife
NCDS1 (15,414)	14,746 14,398 15,414 Varies	Parents Medical examination and tests Medical records School Subject	Interviewed by health visitor Carried out by LA Medical Officer Consulted by LA Medical Officer Questionnaire completed by head teacher & class teacher Tests completed in school
NCDS2 (15,303)	13,879 13,207 14,205 Varies	Parents Medical examination and tests Medical records School Subject	Interviewed by health visitor Carried out by LA Medical Officer Consulted by LA Medical Officer Questionnaire completed by head teacher and class teacher Tests completed by in school

³⁶ This section draws on Ferrie, *Life at 33* and Bynner *et al*, *The design and conduct*.

	13,865	Subject	Questionnaire & essay completed in school
NCDS3 (14,761)	11,691 11,675 12,762 Varies 12,098	Parents Medical examination and tests Medical records School Subject Subject	Interviewed by health visitor Carried out by LA Medical Officer Consulted by LA Medical Officer Questionnaire completed by head teacher and class teacher Tests completed by in school Questionnaire completed in school
EXAMS (14,370)	(14,370)	School attended at 16 and FE colleges when identified	Postal questionnaire
NCDS4 (12,537)	12,537 11,767 12,364	Subject Census 1971 Census 1981	Interviewed by market research interviewer CACI International CACI & UKDA
NCDS5 (11,400)	11,400 11,400 11,400 7,500 2,500 4,000 1,000 5,000	Subject Subject Subject Spouse/Cohabitee Mother of children Children > 4 years Interviewer Mother	Event history self-completions Attitude Self-completions Interviewed by market research interviewer Event history self-completions Mother interviews re: all children Child assessments administered by interviewer Observations re: children < 4 years Questionnaire completed by mother/mother-figure re: all children

5.3.2 The British Cohort Study

The British Cohort Study (BCS also known as the BCS70) is a continuing, multi-disciplinary longitudinal study which takes as its subjects all those in the UK who were born between 5 and 11 April 1970.³⁷ Since its inception the scope of the survey has broadened from a strictly medical focus at birth to encompass physical and educational development at the age of 5, physical, educational and social development at the ages of 10 and 16, and physical, educational, social and economic development at 26 and 29 years.

The BCS began when data were collected about the births and families for a target number of 17,198 babies across the UK in 1970. At this time, the study was named the British Births Survey (BBS) and was sponsored by the National Birthday Trust Fund in association with the Royal College of Obstetricians and Gynaecologists. Since 1970 there have been five attempts to gather information from the full cohort, as shown in Table 7 below. In 1999/2000 an attempt was made to jointly collect information on the subjects of the BCS and NCDS (see section 5.3.1).

³⁷ This section draws on Despotidon and Shepherd, *1970 British Cohort Study* and Bynner *et al*, *The design and conduct*.

Table 7. BCS sweeps and sources of information, 1970-1996

	BBS (1970)	CHES (1975)	CHES (1980)	Youthscan (1986)	BCS70 (1996)
	Birth	5	10	16	26
Information Source	Mother Medical	Parents Test Medical	Parents School Tests Medical Subject	Parents School Tests Medical Subject	Subject
Achieved Sample	16,135	13,135	14,875	11,628	9,003
Per cent Response	98%	80%	93%	72%	56%

Data have been collected from a number of different sources, and in a variety of ways. In the birth survey, information was collected by means of a questionnaire that was completed by the midwife present at the birth, and supplementary information was obtained from clinical records. The five-year and ten-year surveys were carried out by the Department of Child Health, Bristol University and the survey at these times was named the Child Health and Education Study (CHES). In 1975 and 1980 parents of the cohort members were interviewed by Health Visitors, and information was gathered from head and class teachers (who completed questionnaires), the school health service (which carried out medical examinations on each child), and the subjects themselves (who undertook tests of ability). The 1986 survey was carried out by the International Centre for Child Studies and named Youthscan. In this sweep, sixteen separate survey instruments were employed, including parental questionnaires, school class and head teacher questionnaires and medical examinations (including measurement of height, weight and head circumference). The cohort members completed questionnaires, kept two four-day diaries (one for nutrition and one for general activity), and undertook some educational assessments.

In addition to the five major sweeps shown in Table 7, surveys of sub-samples were carried out in 1972, 1973, 1977 and 1991. In 1972 and 1973 the BBS took as its subjects three sub-samples; all twins in the original cohort were included as were low birthweight and post-mature births. A random ten per cent sample of the original cohort acted as a control group. The South-West Region Survey, carried out at the age of three and a half, surveyed ninety five per cent of the cohort members who lived in the south west of England and Glamorgan. These smaller scale surveys were undertaken in order to bridge what was a large gap in child development terms between birth and five years of age (when the next full sweep was planned). In 1977, an attempt was made to obtain details about those who had not responded in the five-year survey. A postal questionnaire was completed by parents of 1,917 of the non-responders. At age 21 a ten per sample survey consisting of 1,600 cases was carried out which focused on adult literacy and numeracy problems in young adults as well as investigating the transition from education to employment.

In both 1975 and 1980, the cohort was augmented by the addition of immigrants to Britain who were born in the target week in 1970. Subjects from Northern Ireland, who had been included in the original 1970 birth survey, were dropped from the study in all subsequent sweeps.

5.3.3 Youth Cohort Study

The Youth Cohort Study (YCS) was designed to monitor the behaviour and decisions of a representative sample of young people aged sixteen and upwards as they make the transition from compulsory education to the labour market, further education or higher education.³⁸ The population studied consists of males and females in England and Wales who had reached minimum school leaving age in the 1998/1999 school year. To be eligible for inclusion they had to be aged 16 on 31 August 1999 (the reference date). To select a sample, DfEE and the Welsh Office wrote to all schools excluding special schools and any institution with less than twenty people, asking them to provide the names of a ten per cent sample of eligible pupils. The sample selection involved taking those born on the 5th, 15th and 25th of the month. The resulting file of 31,424 names and addresses was then reduced by a systematic random selection process in order to produce a file of 25,000 which provided the target sample for Sweep 1.

Sampled respondents were sent an initial pack containing a covering letter, covering notes, questionnaire and business reply envelope through the post. If they did not reply within two weeks they were sent a reminder postcard. If they did not reply two weeks after this they were sent another reminder with a questionnaire in case they had misplaced the original. If they did not reply after a further two weeks then they were sent another reminder pack with a questionnaire. If they still had not replied after this, the contractor attempted to contact the sampled respondent to conduct the interview by telephone.

The sample prepared for Sweep 1 of Cohort 10 was designed to be a representative sample of males and females in England and Wales who had reached minimum school leaving age in the 1996/1997 school year. However, due to differences in response rates by key sub-groups, it is likely that the survey estimates would have been biased had there not been a corrective stage of data weighting: bringing the sample profile into line with the universe for those key variables. For example, while sixty per cent of females responded, only fifty per cent of males did so. Equally, regional response rates varied from sixty per cent for the South West to forty-one per cent for Inner London. Experience has shown that response is often comparatively slow and low for men in general, men and women less qualified academically, and that response varies between different regions and school types.

Given that the one of the main aims of the survey is to provide information on the destinations of young people and that the variables noted as having different levels of response are likely to affect these estimates, then failing to correct for this bias in response would have been likely to result in over estimates of those continuing in full time education at the age of 17. To correct for non-response bias, a weighting strategy was developed, that is, each sample member who returned a completed questionnaire was given a weight which determines the influence that sample member has on the survey estimates. The weighting strategy began with the construction of a population matrix from a small number of variables: sex, type of school attended to age 16, region, and levels of GCSE qualifications attained. The population number falling into each cell of the matrix was estimated from data supplied by DfEE and the National Assembly for Wales. The next stage involved drawing up the sample equivalent of this matrix. YCS is fortunate among surveys in that a number of the variables that are known to be related to non-response to the survey are available on the sample file supplied by DfEE and the National Assembly for Wales and are therefore known for each sample member. Sex, type of school attended to age 16, and region were

³⁸ This section draws on Russell and Phelps, *Youth cohort study*.

available from the sample file. In addition, self reported levels of attainment at GCSE were used in constructing the sample equivalent of the population matrix. Once the weighting algorithm was constructed, a weight was calculated for each sample member. The weights were equal to the population total in the cell divided by the sample total in the cell. The matrices and the weights were then inspected, and, where necessary, groups of cells were collapsed wherever there was likely to be large variability in the weights simply due to random variation, for example because of too small a sample in a cell. In deriving the matrix, the aim was to maximise the variation between cells in the characteristics of sample members, and in response rate, and to minimise variation in the characteristics of sample members within cells. The matrix cells are therefore weighting classes in that all sample members in the same cell will receive the same weight.

5.3.4 The Millennium Cohort Study

The Millennium Cohort Study (MCS) is the most recent major longitudinal survey to be conducted in the UK.³⁹ It was commissioned by the ESRC in 2000,⁴⁰ with support from a consortium of government departments and agencies, with the following stated objectives:

- to chart the initial conditions of social, economic and health advantages and disadvantages facing new children in the 21st century, capturing information that the research community of the future will require;
- to provide a basis for comparing patterns of development with the preceding cohorts (NCDS, BCS);
- to collect information on previously neglected topics, such as fathers' involvement in children's care and development;
- to focus on parents as the most immediate elements of the children's 'background', charting their experience as mothers and fathers of newborn babies in the year 2000, recording how they (and any other children in the family) adapted to the newcomer, and what their aspirations for her/his future may be;
- to emphasize intergenerational links including those back to the parents' own childhood;
- to investigate the wider social ecology of the family, including social networks, civic engagement and community facilities and services, splicing in geo-coded data when available.

Additional objectives subsequently included were:

- to provide control cases for the national evaluation of 'Sure Start' (a government programme intended to alleviate child poverty and social exclusion);
- to provide samples of adequate size to analyse and compare the smaller countries of the UK.

The MCS is managed by the Centre for Longitudinal Studies at the Institute of Education, University of London, with fieldwork being conducted by the National Centre for Social Research and the Northern Ireland Statistics and Research Agency's Central Survey Unit. Fieldwork was conducted between June 2001 and September 2002 in England and Wales, and between September 2001 and January 2003 in Scotland and Northern Ireland.

³⁹ See Shepard *et al*, *Millennium Cohort Study* and

⁴⁰ ESRC award number: H567255003.

Those targeted for interview were parents of living babies born between 1st September 2000 to 31st August 2001 in England and Wales, and between 22nd November 2000 and 11th January 2002 in Scotland and Northern Ireland. The sample population of eligible live births was selected from a random sample of electoral wards, disproportionately stratified to ensure adequate representation of all four UK countries, deprived areas and those with high concentrations of Black and Asian families. Thus, children with sample birth dates were eligible for the survey if they were living in one of approximately 400 electoral wards across the whole of the UK when they were nine months old. This produced a target of 20,646 children, of whom information was obtained from 18,819.

In 'locating' the target children help was offered by the then Department of Social Security⁴¹ to make Child Benefit records available to find the children. This helped to address potential problems of representativeness, which may have resulted from using birth registration records in isolation, particularly in relation to socially excluded groups, including those with poor literacy or poor grasp of English. Because the Child Benefit records do not reveal all families who may have moved into the sample wards as the child approached nine months, an attempt was made to find movers-in with the help of Health Visitors. These were asked to see if families moving into survey wards were willing to be recruited and were also asked to inform and reassure other families, who might have received information about the survey.

The second wave fieldwork is planned to start in September 2003, when the children will be aged three years.

5.4 International historical longitudinal studies

5.4.1 The Historical Sample of the Population of the Netherlands

Based at the International Institute of Social History, the Historical Sample of the Netherlands (HSN) has its origins in 1987 when members from several Dutch universities formed a working group to examine the feasibility of constructing a historical longitudinal database for the Netherlands.⁴² A pilot project was carried out on the province of Utrecht since, of all Dutch provinces, this was deemed as having the closest demographic profile to the Netherlands as a whole. Following this successful feasibility study, the Utrecht project was used as a prototype for the main study, extended to the whole of the Netherlands.

The sample is based on two sources: the Hofstee database, which covers the years 1811 to 1850; and the Historical Ecological database, which covers the years 1851 to 1922. It is important to realise, in comparing the Dutch case with the proposed VPS, that these sources are in effect population registers. Thus they are themselves longitudinal records, and therefore the basic problem of tracing individuals overtime, and linking administratively different sources, is largely overcome in the Dutch case.

The sample was based on the official records of the number of births each year in each municipality. These births were stratified into cohorts of ten years, with the exception of the first cohort (1812-1822), which consisted of eleven years. Records were then stratified by the place of residence. However, the Hofstee and Historical Ecological databases do not overlap exactly with the number of birth certificates. This is because, using the example of Utrecht, more children were

⁴¹ Now Department for Work and Pensions.

⁴² His account of the HSN draws on Mandemakers, 'Historical sample of the Netherlands (HSN). Background'; Mandemakers, 'The Netherlands' and Mandemakers, 'The historical sample of the Netherlands'.

born and given birth certificates than were officially counted as newborn children in that city. Mothers came to the city to give birth, and then either returned home or disappeared from the records. To compensate for this anomaly the official number of births was enlarged by ten per cent. This gave every birth certificate an equal chance of being sampled. However, the enlargement to one hundred and ten per cent of the official number of births meant that many sampling numbers did not correspond to a birth certificate. These sampling numbers were disregarded.

Employing this method the resulting initial basic sample frame of the HSN consisted of a half per cent of all birth certificates. Yet this led to inequality in the numbers per cohort who survived to age twenty. This problem was revealed by the pilot project, and it was decided to employ a more sophisticated sampling technique, varying sample ratios differentially according to period. The ratios used were:

1812-1872	0.75%
1873-1902	0.50%
1903-1922	0.25%

The resulting HNS contains about 80,000 main entries. Loss of data from observation overtime stands at about forty per cent, due mainly to early deaths and migration.

The structure of the HSN consists of five normalised database files. These files consist of births/deaths, marriages, marriage witnesses, occupations and locations. The first file contains the birth records of the original person and a number of variables from the death records. Most information concerns persons who died young in the community they were born in. In addition to this, the dates of last observations, gathered from personal cards, population registers and marriage certificates, were included.

The second file contains variables on married couples obtained from marriage registers. This includes details of second and third marriages. The third file contains witnesses to marriage records organised by the characteristics of the married couple. This includes up to four persons per marriage record. The fourth file contains details of specific occupations together with their classification codes. The last file contains tables of specific names of municipalities and other types of location, along with classification codes and geographical coordinates.

The HSN software consists of six clusters. Four of these are for the entry, administration and control of the data. One contains the database and one contains the cluster release. This enables data release and data description. The software used has undergone a continuous process of refinement and checking.

The second phase, which is ongoing, concerns the gathering of more material from marriage certificates, population register entries and death certificates for those dying before 1940. There are further ongoing projects being carried out in collaboration with other institutions. These often include additional sampling being carried out to secure sufficient records for specific research. This additional sampling is not necessarily based on birth certificates.

A recent development with the HSN has been a collaboration with the Netherlands Historical Data Archive (NHDA),⁴³ who since 1997 have been working in partnership with Statistics Netherlands, and more recently with the Historical

⁴³ A section within the Netherlands Institute for Scientific Information Services (NIWI).

Databank of Dutch Municipalities, to digitise the published census reports, and to convert the tabular aggregate statistics included in these volumes to database format. This work is particularly important given the fact that no individual level census data survive for the Netherlands prior to 1960. Linking the data from population censuses and other municipal data with the HSN will provide a huge amount of contextual information within which the HSN database can be explored, creating new opportunities for multi-level and cross-level analysis.

5.4.2 The linking of historic census samples for the USA

The Minnesota Population Center has submitted a bid to the NIH, which is shortly to be funded, to create linked representative samples of individuals and family groups from the existing national samples of the censuses of 1860, 1870, 1900, and 1910 to the 1880 census, which exists in computerised form at the one hundred per cent level.⁴⁴ Using new record-linkage technology, it is proposed to construct linked samples covering pairs of census years: 1860-1880, 1870-1880, 1880-1900, and 1880-1910. Each of the samples will be independent, but taken together they will provide a rich longitudinal source for the nineteenth and early twentieth centuries.

Using the complete 1880 database as a base, the project will utilise recent innovations in data-mining and automated linkages techniques to link the 1880 population with each of the other existing census samples including nominal information. However, unlike most linking projects which focus on maximising the number of accurate links, this project will be designed to maximise the representativeness of the linked cases, paying close attention to sources of selection bias, and ignoring much of the information routinely used by other record linkage procedures. For example, record linkage algorithms ordinarily make use of place of residence as a linking variable. This greatly increases the potential for reliable links: if an individual is identified in the 1870 sample who partially matches the name and other characteristics of a person in the 1880 database, the confidence that the two records refer to the same person would be improved if it was also known that they both reside in Poughkeepsie. Yet if place of current residence is used in the linking algorithm, it is argued, this will inevitably bias the sample in favour of non-migrants. Likewise, if spouse's name is used in the algorithm it will bias the sample in favour of those who remain married, and if occupation is used it will favour cases with low social mobility.

Three categories of linked samples, each with a different universe are planned:

- all males;
- females who do not marry in the census interval;
- married couples.

Although none of these groups is representative of the entire population. The project's goal will be to make each category representative of its defined universe. The male individual sample will be general purpose, useful for studying economic and geographic mobility, transitions to adulthood, changes in family composition, and retirement. The female sample will be useful for studying many of the same topics, but will apply to the subset of women who do not change their surname between censuses, and therefore will be inappropriate for some topics. The married-couple samples will offer the greatest reliability, since it will allow linking on characteristics of both husband and wife, and will be especially useful for topics relating to fertility, child mortality, and age of leaving home. Because it

⁴⁴ This section of the report is based on the application to the NIH, kindly supplied by the PI for the project, Professor Steven Ruggles.

is restricted to the continuously married population, however, it will be less useful for population-wide generalisations about social and geographic mobility. Although the linking 'unit' will be individuals or couples, information will also be captured on the characteristics of all other co resident household members.

For each sample, the project will start by identifying a subset of individuals in the existing census one-percent samples (1860, 1870, 1900, or 1910). A search for these will then be undertaken on the complete 1880 census database. Three linked samples for each pair of census years will be created, for a total of twelve samples. Half of the samples will use forward links (1860 and 1870 to 1880) and half backward links (1900 and 1910 to 1880). Forward-linked samples are more challenging than the backward-linked ones because mortality and emigration substantially reduce the potential for links.

The linking algorithm to be designed and implemented will rely exclusively on characteristics that should not change over time. The linking characteristics that are planned to be used for each census year are given in Table 8. For the married-couple samples, these characteristics are available for both husband and wife.

Table 8. Variables available for record linkage, 1860-1910

1860 to 1880	1870 to 1880	1900-1910 to 1880
First name	First name	First name
Last name	Last name	Last name
Birth year	Birth year	Birth year
Sex	Sex	Sex
Race	Race	Race
State or country of birth	State or country of birth	State or country of birth
	Father foreign born	Father's state or country of birth
	Mother foreign born	Mother's state or country of birth

The chief problem posed by this approach is that this limited set of variables is insufficient to identify individuals uniquely. For example, the 1880 census includes seventeen white men aged thirty-three who were named John Smith and born in New York State. This is, of course, the worst-case scenario, since John Smith was the most common male name and New York was the largest state. Nationally, it is expected that seventy-five per cent of the population can be uniquely identified by the limited variable set available in 1860. For the married-couple universe, virtually every case would be uniquely identified.

In practice, however, those numbers are optimistic. Because of errors in enumeration and transcription, a high proportion of matches are imperfect: linking must be carried out on a probabilistic basis, allowing for imperfect correspondence of names and ages. Allowing for such near matches, the proportion of uniquely identified individuals would decline significantly. To reduce the potential for ambiguity, the project will follow the precedent of Ferrie⁴⁵ and eliminate the most common names. A set of names will be defined that reduces the potential for ambiguous links but also maximises representativeness with respect to ethnicity, occupational status, and other characteristics.

⁴⁵ See Ferrie, 'A new sample of males' and Ferrie, 'How ya gonna keep 'em down on the farm'.

Once the project has established the links, it will add the omitted 1880 variables covering health, education, unemployment, and other characteristics for all persons residing in a household containing linked individuals. There is insufficient information at present to estimate the total size of the linked samples, but it is estimated that approximately 600,000 cases will be created across all linked samples. The project will also provide a confidence score for each linked case, indicating the strength of the similarity measures between the two cases, and construct weights for the linked samples that maximise the representativeness of the linked cases with respect to occupation, birthplace, parental birthplaces, age, sex, race, marital status, literacy, state of residence, urban residence, and school attendance.

Although this projects differs in approach and outcomes from that proposed for the VPS, it could provide opportunities for comparative research on nineteenth-century American and British populations. The software that will be produced for automatic linking could also have implications for the VPS, and taken together, each can provide useful and instructive information on the potential bias in the other and the measures needed to correct for this.

5.4.3 Swedish demographic databases

The Demographic Database (DDB), based at the Centre for Population Studies, Umeå University, Sweden, began as a temporary employment project in 1973 under the auspices of the Swedish National Archives.⁴⁶ Its primary aim was to computerise historic parish records and make them available for research.

The DDB consists of data from historical parish registers and parish statistics from Sweden, mainly relating to the nineteenth century, but also containing earlier records. The primary sources were catechetical lists, ministerial lists of births marriages and deaths and migration lists. The initial study was limited to seven individual parishes, it was then extended to Skellefteå and Sundvall and their surrounding areas.

The DDB is restricted to a fairly small region and does not employ any form of recognised sampling technique. As such is a selection of longitudinal records, rather than a sample. However, the project has recently developed a suite of computer programs for automated record linkage, and these may be of relevance for the proposed VPS.

In addition to the DDB another important and incredibly rich collection of historical longitudinal data in Sweden is the Stockholm Historical Database.⁴⁷ This is an ongoing project which aims at creating a complete demographic database of all residents of Stockholm between 1878 and 1926, based on the special population register of the Roteman Archive. Information is currently available for over 250,000 individuals. The database provides details of occupation, poverty, medical provision, crime, mortality, education, migration and mobility, family and household structure. One particularly interesting feature of the Stockholm Historical Database is that the underlying individual and household records have been linked to an historic Geographical Information System (GIS) of Stockholm allowing the data to be mapped and visualised spatially.

Another important group in Sweden is that undertaking research in historical demography based at the University of Lund. Whilst working primarily on parish registers and records, this group has developed computer programs for

⁴⁶ See Stenflo and Sundin, 'Using a large historical database'.

⁴⁷ See Fogelvik, 'The Stockholm Historical Database at work'.

automated record linkage and has an established track record in the analysis of historical longitudinal data.⁴⁸

5.4.4 Programme de Recherché en Démographie Historique

The Programme de Recherché en Démographie Historique, at the Université de Montréal is a project of long-standing which aims at reconstituting the population of historic Quebec from the period dating from the beginning of French colonization. This reconstruction takes the form of a computerised register of population that contains the biographical files of all the individuals of European stock who lived in the valley of the St. Lawrence. It consists of data from baptismal, marriage and burial certificates from 153 parishes, missions and institutions.

Each personal file specifies the date and place of birth, as well as details of marriages and deaths, and the subsidiary and marital bonds maintained with other individuals. The basic information is supplemented by socio-demographic characteristics such as occupation, literacy, place of residence and, in the case of immigrants, the place of origin, as drawn from the historic source records..

A computer-automated procedure has been used to identify all potential links and choose the most likely solutions on the basis of name frequencies and demographic criteria. Unambiguous links were resolved automatically, while others are resolved manually.⁴⁹

⁴⁸ See, for example, Bengtsson and Lundh, *Evaluation*.

⁴⁹ Bouchard, 'Current issues'.

6 A potential methodology for the VPS

From the previous section it is clear that there is no such thing as a perfect longitudinal study. Each potentially comes with elements of bias, and moreover, all suffer from problems of non-response where it has not proved possible to track target individuals.

Clearly there are two basic types of longitudinal studies. First, cohort studies in which individuals from a particular age group are the initial sampling target and are subsequently traced over time. Second, panel studies in which residential households are the initial sampling target and from which one or more selected members are traced over time. Obviously, these two basic approaches can overlap to some extent, since a common feature of most longitudinal studies is to collect information not only on the original sample or target members (OSMs) but also on those with whom the OSMs resided at certain points. These individuals can be viewed as attributes to an OSM or temporary sample members (TSMs). Thus, it is possible to use some cohort studies, such as the LS, as a quasi-household panel survey.

6.1 Selecting the Original Sample Members

Assuming that the VPS has a starting point in 1851, given that this is the first census date at which sufficient information is available, it would be possible, for example to draw a national sample from all those born over the course of the year prior to the taking of the 1851 census from the indices to the civil registers, define these as OSMs and track them over the period 1851 to 1901. Likewise, it would be possible to draw a sample of households from the 1851 CEBs, define one or more of the selected household members as OSMs, and track these. A third possibility might be to select all those in 1851 who have one of a set of randomly selected surnames. However, each of these options would result in significant costs first to establish the sample members and then create a database of the OSMs and associated TSMs as recorded in 1851.

A more pragmatic and cost effective solution to this problem would be to select OSMs from the already existing national sample of the 1851 census, a project into which the ESRC has already invested relatively heavily. Yet even with this solution a fundamental question still remains: who is selected as an OSM? There would seem to be three basic options to choose from:

- select everyone recorded in the 1851 sample as an OSM;
- create a sub-sample of the 1851 sample from which OSMs are drawn;
- select OSMs from a specified age-cohort from within the 1851 sample.

Each basic option comes with its advantages and disadvantages. Given that the existing 1851 database is a nationally representative random stratified sample the first option probably poses fewest methodological problems, but it also undoubtedly creates the greatest workload. The 1851 sample consists of 393,638 individuals enumerated across 82,470 households and 323 institutions. If all were to be designated as OSMs this would be significantly higher than an existing longitudinal study for a modern day population, by a factor some twenty times. It also means that by 1901 one might expect the number of OSMs to probably have risen to 750,000 (adopting, for example, the rules applied in the BHPS), with as many as 3,000,000 attribute members or TSMs being observed between 1851 and 1901. There are, however, clear advantages of having large numbers in any longitudinal study as this will facilitate the ability to examine potentially small

sub-groups of the population, for example in following the experiences of a relatively small occupational group, examining a specific demographic sub-group such as the 'extreme' old, or conducting analyses according to regional breakdowns. It is also the case that one cannot directly compare the resource needed for tracing and moreover capturing the data necessary for modern day longitudinal studies with historical longitudinal studies. For the latter, the task of tracing and linking may be more time consuming (as discussed later), but the job of compiling the data for successive waves is far less resource intensive. This is due to three basic reasons: first, there is less information to collect; second, a significant amount already (or will) exist in computerised form; third, and most importantly, time is not spent in lengthy interviews or similar enquiries. It is hard to provide any concrete information without a full feasibility study, but it may well be that it is ten or even twenty times quicker to construct a new wave for an historic longitudinal study than it is for that on a modern day population.

If size is deemed to be a problem, then obviously the scale of the VPS can be reduced by drawing a sub-sample from within the existing 1851 sample according to a specified sampling fraction. Alternatively (or in addition) the 1851 sample could be further stratified in order to identify OSMs. One strategy might be to randomly select one individual from each household as an OSM.⁵⁰ This may not seem to be a radical shift from the option of selecting everyone as an OSM, since by definition, still all the 393,638 individuals in the 1851 sample living in or attached to households would be an OSM, would become an OSM (such as children of a person selected as the OSM), or would be a TSM. Yet in considering this option, it must be realised that the structure and composition of households in 1851 was significantly different from those of modern day populations in one important regard, in that they contained a large number of non-core family members. Thus, taking the person enumerated as the 'head of household' as the reference person, in the 1851 sample the following pattern emerges:

- 26,842 individuals (6.74%) were non-nuclear family relatives;
- 17,898 individuals (4.49%) were inmates (lodgers and boarders));
- 25,277 individuals (6.34%) were servants;
- and 8,237 individuals (2.07%) were miscellaneous 'visitors'.

Thus, some twenty per cent of the population in 1851 were living outside of what today might be thought of as 'standard' nuclear families. This also means that selecting an individual at random from each 1851 sample household as the OSM, for example, up to twenty per cent of the initial census population would be dropped by 1861 as there would only have TSM status in 1851.

In relation to the possibility of defining OSMs by sub-sampling the initial 1851 sample, it is important to remember that the 1851 sample is primarily a sample of EDs from which all enumerated individuals and households are included in the sample. This geographic clustering potentially brings its own, although not insurmountable, problems.⁵¹ Yet equally, those consulted about the plans for the VPS currently working on present day longitudinal studies, were excited by the prospect of including what might be termed a community profile in the VPS. This would allow the data subjects to be analysed within a community context through the examination of the data in micro geographical or environmental contexts, a feature which is not possible with modern day longitudinal studies. Retaining a

⁵⁰ However, the result would not be a pure random sample, as individuals initially resident within small households would by definition be over-represented.

⁵¹ For example, if every one in the 1851 sample is selected as an OSM, those unmarried in the original sample will be more likely to select existing OSMs drawn from within the same geographic cluster, than non-OSMs as their future spouses.

community element to the VPS would also importantly allow the creation of associated data collections which could be utilised alongside the main VPS. This is a feature that particularly interests TNA and other potential partners to the project, such as the AHRB (see section 8.4).

The third basic option identified above is to select an age cohort from the 1851 sample and define these as OSMs. A possibility within this strategy would be to select all those aged less than five, for example. This would provide the following numbers as a cohort group:

Age	n.	% of total population
<1	13,642	3.42
1	10,116	2.54
2	10,504	2.64
3	9,636	2.42
4	9,850	2.47
<5	53,748	13.49

However, creating an age cohort from the 1851 sample is not so straightforward. Realistically if a cohort approach were to be taken one would wish to construct a birth cohort rather than a cross-sectional period cohort. The key point is that of those born in the period 1846 to 1851, say, several will have died before being at risk to be recorded in the 1851 census. The larger the cohort in terms of age, the larger the number who will have died prior to the census point. Thus, ideally, any children of parents selected from the 1851 sample who were born and died before the 1851 sample would have to be traced from the civil registers and 'added' to the sample. Although not impossible, this could be a time consuming task, with the result that if a cohort approach were to be taken, it may prove practical to do this only for those born, say, within the twelve-month period prior to the census date. Given the mortality rates of the mid-nineteenth century, for the whole 1851 sample this would mean adding in some 2,050 infant deaths if the birth cohort April 1850 to April 1851 were to be taken, and 12,700 births and subsequent deaths if the birth cohort were to be extended to April 1846 to April 1851.⁵²

At this stage it would be unwise to suggest, categorically, what method is adopted to identify and select OSMs for the VPS. Indeed, an optimum approach might be to adopt a flexible strategy in which a variety of cohort, panel and 'community' methodologies were implemented. This could produce a number of separate 'views' of the VPS database which could be used differentially according to the research questions to be addressed. Clearly it would be important objective of the proposed pilot project (see section 9) to examine the possibilities, problems and resource implications of each of these basic approaches.

6.2 Who remains in observation?

Regardless of which strategy or strategies might be adopted for the VPS, a related issue that will need to be resolved is the question of who remains in observation, since this will potentially greatly impact on the number of individuals and families that will be at risk to trace, and therefore the overall workload.

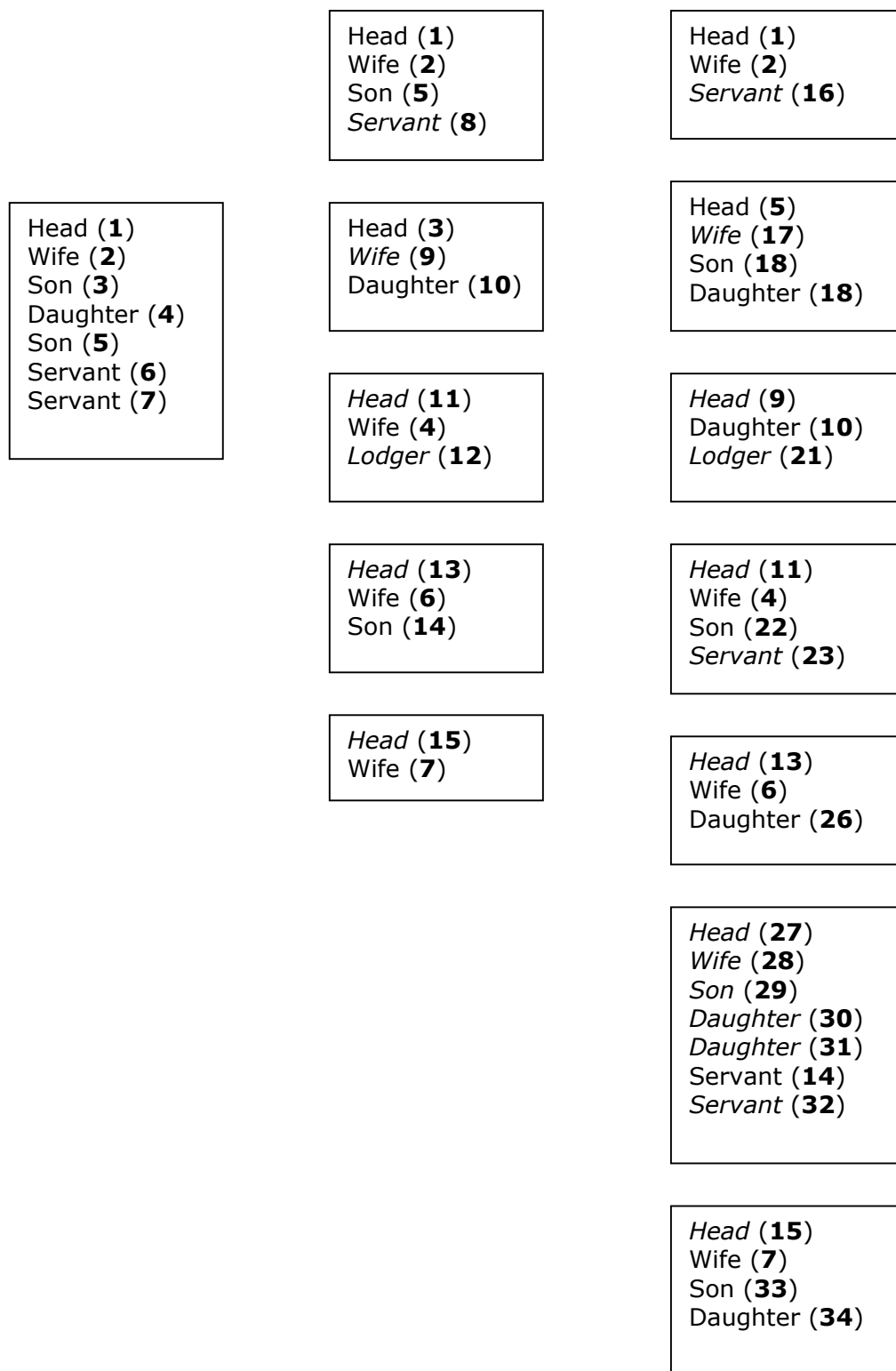
This is illustrated in the two hypothetical scenarios shown below.

⁵² These estimates are based on the mortality levels given in the Third English Life Table covering the period 1838-54, as reported in Wrigley and Schofield, *Population history*, 709.

Scenario I

	Head (1) Wife (2) Son (5) Servant (8)	Head (1) Wife (2) Servant (16)
Head (1) Wife (2) Son (3) Daughter (4) Son (5) Servant (6) Servant (7)	Head (3) Wife (9) Daughter (10)	Head (5) Wife (17) Son (18) Daughter (18)
	Head (11) Wife (4) Lodger (12)	Head (19) Wife (8) Daughter (20)
	Head (13) Wife (6) Son (14)	Head (9) Daughter (10) Lodger (21)
	Head (15) Wife (7)	Head (11) Wife (4) Son (22) Servant (23)
		Head (24) Lodger (25) Lodger (12)
		Head (13) Wife (6) Daughter (26)
		Head (27) Wife (28) Son (29) Daughter (30) Daughter (31) Servant (14) Servant (32)
		Head (15) Wife (7) Son (33) Daughter (34)

Scenario II



In both of these diagrams the three columns of boxes moving left to right represent the three census years 1851-71. The numbers in bracket after the hypothetical household members are included as an aid so that 'individuals' can be tracked over time. For example, the hypothetical Head numbered (1) is the same individual recorded in the three census years, living with his wife (2) the

whole time, but gradually witnessing the departure of his children from the parental home.

Scenario I represents an extreme case in which all of the household members in a household in the 1851 sample are treated as OSMs and are tracked over time, but also in which all other individuals they reside with (at any point in time) are also automatically assigned OSM status and also subsequently tracked. In this case the size of the population under observation grows from 7 to 15 to 34, with the number of households recorded increasing from 1 to 5 to 9.

Scenario II represents a situation similar to that implemented by the BHPS. In this case, as with scenario I, all of the household members in a household in the 1851 sample are treated as OSMs and are tracked over time. However OSM status (with subsequent tracking) is only passed onto the offspring of OSMs. All other individuals recorded as co-residents of OSMs are assigned TSM status only and subsequently dropped from observation if they no longer appear in the household of an OSM (TSMs are indicated in the second diagram by italics). In this case the size of the population under observation grows from 7 to 15 to 28, of which 14 are TSMs and viable to be dropped, while the number of households recorded increasing from 1 to 5 to 7.

From these brief illustrations it can be seen that it is probably neither practical nor desirable to implement observation rules such as those sketched in Scenario I. In the case of the VPS it would seem logical and desirable to use observation rules based on those of the BHPS. This will ensure that the target population at risk to be traced does not 'mushroom' to unmanageable proportions; will equally ensure that sufficient attribute data is collected (based on TSMs); and importantly, will allow comparisons to be drawn between the two studies.

6.3 Linkage

In the case of the VPS the tracing of individuals over time probably presents a much more problematic task than is the case for most modern day longitudinal studies. As has already been discussed in section 3.3 of this report, academic researchers are no strangers to linking individuals successively across historical documents and have developed a series of automated computer programs to aid the process.⁵³ However, nothing has been attempted on the scale of the proposed VPS, and as mentioned previously, academic historians have made very limited use of the civil registers in record linkage exercises due to the current access arrangements to these data.⁵⁴

However, unlike academic historians, for genealogists and family historians the tracing of individuals across *both* census documents and civil registers is a stock-in-trade. Yet their approach in doing this, from necessity, is quite different from

⁵³ See, for example, Adman *et al* 'Computer-assisted record linkage'; Bloothoof, 'Assessment'; Bouchard, 'The processing'; De Brou and Olsen, 'The Guth algorithm'; Guth, 'Surname spelling'; Harvey and Green, 'Record linkage'; Harvey *et al*, 'Record linkage'; Hershbery *et al*, 'Record linkage'; Katz and Tiller, 'Record linkage'; Kitts *et al*, *Viana do Castelo*; Lait and Randell, 'An assessment'; Miller and Thorvaldsen, 'Beyond record linkage'; Mineau *et al*, 'Description and evaluation'; Morton, 'Presenting the self'; Tilley and French, 'Record linkage'; Wrigley, *Identifying people*.

⁵⁴ It is the case that access to civil registration data has been more flexible in Scotland, with the result that a number of studies have made use of the Scottish material. See, for example, Anderson *et al*, *Demographic change*. In the case of England and Wales one notable exception is the project that linked various historical sources for mid-nineteenth-century Ashford, for which the local Registrar gave permission for the research team to consult the civil registers held locally. See, Drake, 'Ashford'.

the academic historians utilising computerised linkage techniques to establish links. Genealogists link manually, proceeding sequentially (sometimes backwards in time) from one source to another. The VPS offers potential to combine the two approaches.

The availability of, and access to, the various indices that have been made, or are planned, is fundamental for a project the size of the proposed VPS. Complete indices for the censuses and civil registers will allow researchers to search for individuals at a national rather than local level, and thus address the issue of migration, a basic problem with all local-based linkage exercises. However, where these tools currently exist, for example, with the 1901 census, the way they are used by researchers (mainly genealogists) is to conduct a search for one individual at a time. Indeed, this is exactly how the various on-line search engines for indices like the 1901 census have been designed. Although a great advance on the finding aids available only a few years ago, trying to trace one individual at a time through the various sources would prove far too time consuming for the VPS. Instead, it is planned to harness the databases behind the existing genealogy-orientated search engines (as well as those in the process of being created, or scheduled for creation).

It is too early to say at this stage what access the VPS might have to the civil registration indices and certificate database and in what form such access might take. However, initial discussions with ONS suggest that they would probably be well disposed to providing support and assistance to the project. The FreeBMD initiative, who have already created a computerised index covering 58 per cent of the civil registration events (some 51.3 million entries so far) and who expect to complete the index within the next year, have already agreed in principal to provide the VPS with copies of their entire database, free of charge. Likewise TNA are committed to providing the VPS with privileged access to all future census indices and related databases that it creates. Lastly, discussions with QinetiQ, who produced the existing 1901 census index and database in association with TNA, suggest that they may be willing to provide the VPS with special access to the material they hold.⁵⁵

It has not proved possible within the short timescale of preparing this report to ascertain a definite statement of the position of the GRO(S) regarding the proposed VPS. In Scotland, as reported earlier, the civil registers have already been computerised, and it is estimated that all of the historic censuses (1861-1901) will be computerised by the end of 2004. Although it is not known what special access the VPS may be granted to these resources, if any, encouragingly one project recently funded by the ESRC linking data for Skye and Kilmarnock, has managed to secure special access to the computerised data held by GRO(S).⁵⁶ Equally, other researchers in Scotland have reported on the willingness of GRO(S) to work in co-operation with academic projects.

In addition to the 1851 national census sample and the 1881 census database covering the entire British Isles, both of which are already available for academic research, unprecedented opportunities are opening up for tracking named individuals over the course of the second half of the nineteenth century. Building on these opportunities, it would be possible to create a large database incorporating some 257 million records on which automatic record linkage techniques could be applied. With advanced record linkage and data mining techniques, such as those being developed at the Minnesota Population Center,

⁵⁵ QinetiQ would probably wish to negotiate payment for this access.

⁵⁶ For which a payment of £1,600 has been made.

the task of tracing those recorded in the 1851 national sample, previously unthinkable, could soon become a possibility.

But, realistically, automated linking and data mining will not solve all the problems. Manual intervention would be needed for both checking the 'quality' of the computer generated links, and resolving multiple and ambiguous links, a particular problem in parts of the country where surname pools are shallow, with many people sharing a relative small number of surnames.⁵⁷ In this respect, automated record linkage, even on a database of 257 million person records, might perhaps best be seen as providing solutions for unambiguous links, and providing pointers or signposts to the resolution of ambiguous links.

Because of this need for manual intervention, verification, and possibly further data collection, it is proposed that the VPS draw on volunteers from the genealogy, family history and local history communities. This is a large and growing body of individuals: the Federation of Family Societies has a membership of some 180,000. Initial discussions with representatives from these communities has suggested that there are many who would offer their time (free of charge) to assist the work of the VPS. Equally, TNA who are the main service providers for family historians, are willing to promote the VPS within these communities.⁵⁸ Combining the new opportunities opening up to undertake automated record linkage on those living in the second half of the nineteenth century, with the willingness and enthusiasm of the large and highly active family and local historian communities makes the VPS a realistic possibility.

However, as with any exercise involving the tracing of individuals over time, both present and past, it will not be possible to create a one hundred per cent fully linked sample. Although it is expected that linkage rates will be high, some sample members will remain untraceable and bias will result, as is the case in all modern day longitudinal studies. Thus it will be most difficult to trace the very mobile and disadvantaged groups such as the poor. In this respect, measuring emigration, whilst possible, will be problematic since it will be hard to distinguish between those who cannot be traced due to emigration from Britain and those who cannot be traced due to the inadequacy of the sources.

Equally, for the VPS it is intended to make linking decisions purely on the basis of census and civil registration information alone. Although, arguably, linkage could be improved by introducing additional sources, unless this is done systematically it would result in creating bias within the sample as this would impact on an individual's chances of being traced successfully overtime. Thus, the intention is to restrict the VPS to the censuses and civil registers alone. The downside of this strategy is that the VPS will contain limited information on the process of change from one census to another. For instance, as in most studies, the details of migratory moves that occurred between censuses may go unrecorded, if not indicated by any intervening civil registration. Discussing this problem with representatives of the genealogical and local history communities suggests that some volunteers may be interested in linking other sources to the 'core' VPS for selected families and places. Although this subset of the VPS, if created, could not

⁵⁷ However, as noted above, the Minnesota project plans to adopt a fully automated solution in order to attempt to minimise linking bias.

⁵⁸ The idea to call upon volunteer researchers from the local and family history communities is not new. Although the projects differed both in terms of the sources used and approaches adopted, volunteers were used to remarkable effect in the case of the ESRC-supported work of the Cambridge Group for the History of Population and Social Structure, and in the ESRC-funded research undertaken by Colin Pooley and Jean Turnbull on migration and mobility in Britain between 1750 and 1950 (ESRC award number R000234638). See Pooley and Turnbull, *Migration and mobility*, 31-8.

be seen as statistically representative at a national level, it may prove useful in providing contextual information on topics that the VPS cannot address directly, as well as providing valuable information on how to interpret some of the results to come from the VPS.

6.4 Refreshment

Regardless of how good the linking on the VPS might be, the issue of refreshment will still need to be addressed. While the 1851 census sample is representative of Britain at that point in time, even if everyone recorded in the sample was traced to either their death or the 1901 census with total accuracy, the resulting population in observation at 1901 would no longer be representative of Britain as a whole. For example, by definition all those remaining in the sample in 1901 would have to be able to trace their roots to residence in Britain in 1851. Thus, none of the immigrant groups that entered Britain, such as the Jews who flooded into the East End of London between 1880 and 1900, would be present in the sample.

This type of problem is overcome in some longitudinal studies by applying a refreshment strategy in which new targets (individuals or households) are added to the initial sample. These are selected in order to make the sample representative in the light of changing demographic, socio-economic or geographical structures. It will prove to be desirable to apply a refreshment strategy to the VPS, but it is too early at this stage to say what such a strategy should be. Much will depend on the extent to which the geographical clusters which make up the 1851 national sample remain representative of Britain over time, and it is no quick and easy task to calculate this.

6.5 Project organisation

A project like the proposed VPS will not just happen, it will need accomplished project management and organisation to succeed. This is particularly the case if a large and geographically diverse group of volunteer researchers is to be used. For this reason it is suggested that the VPS will probably need the ESRC (maybe with others) to fund a central VPS Research Unit that will have overall responsibility for the management and co-ordination of the project. This body would:

- liaise, communicate and report to stakeholders;
- co-ordinate a Management Board;
- co-ordinate user community input;
- recruit, train and encourage volunteer researchers;
- verify and quality control work of volunteers;
- design and manage the central VPS linkage database;
- develop automated record linkage software;
- design and populate the core and ancillary VPS databases;
- code and standardise the core and ancillary VPS databases to maximise compatibility with other longitudinal studies;
- produce guides and documentation for the VPS;
- provide training in the use of the VPS;
- promote the use of the VPS.

Another important function that the VPS Research Unit would need to undertake is the actual checking and linking of data. Although the intention would be to recruit as many volunteer researchers as possible, inevitably some individuals, households or even places in the VPS, will not have people volunteering to work on them. As a result, the Research Unit will need to have researchers in house

who can take on the work of reconstituting the 'orphans' left unadopted by the volunteers.

A further role for the Research Unit would be to develop 'intelligent' and user-friendly software that could be disseminated to the volunteers so that they can easily check links, resolve multiple and ambiguous links and enter in new data where necessary. Such software could be CD or internet based (according to volunteer needs) and could incorporate, like CAPI software used in present day surveys, facilities for automated coding and consistency checking.

While it is probably more efficient and cost effective to centralise most of the activities of the proposed Research Unit, it may prove desirable for some tasks to be undertaken by other specialised units working in a distributed infrastructure.

Once the VPS has been completed, it will be passed to ESDS Longitudinal and/or AHDS History for preservation, dissemination and support. It may prove desirable to produce different releases of the data, say 1851-61, followed by 1851-61-71, and so on, but this would be very dependent on the uptake from and progress of the volunteer researchers.

7 Related projects

7.1 The Historical Population Reports Collection

The JISC has recently agreed to fund a major digitisation project to be run by AHDS History at the UKDA, University of Essex. This three-year project, due to start in early 2004, will create a digital version of the complete published population statistics for the United Kingdom for the period 1801 to 1931. The key deliverable for the project will be a web-based user interface allowing browsing, searching, viewing and downloading of approximately 190,000 images of historical population reports. The interface will also allow the viewing and downloading of machine-readable versions of a number of statistical tables contained within the reports.

The project also aims to bring together previously created and newly created resources relating to the population of the nineteenth and early twentieth century. This will be achieved by integrating images of published census and registration statistics, resources which are increasingly difficult to access, along with files containing machine-readable versions of the tables contained within these reports, thus allowing their immediate use by interested scholars.

The key sources which will be digitised and disseminated are the published reports arising from the census and the annual and decennial reports of the Registrars-General of England and Wales, Scotland and Ireland. In addition, a selection of associated manuscript documents in TNA, which give insight into the administrative procedure followed in the taking of the census and the collection of registration material, will be digitised.

In summary, the key objectives of the project are:

- capturing the images from approximately 187,000 pages of population reports from a variety of sources;
- capturing the images from approximately 5,000 pages of unpublished archival material at TNA;
- capturing the text from approximately 50,000 pages of population reports at searchable quality;
- integrating previously existing machine-readable versions of statistical tables from over 200 tables within the database;
- creating approximately 200 further machine-readable versions of statistical tables and subsequently integrate these within the database;
- development of a web-based user interface for browsing, searching, viewing and downloading images from the published reports and all ancillary material.

The majority of image capture will take place at the Centre for Digitisation and Data Analysis at Queen's University, Belfast, with some additional image capture taking place at TNA.

As with the collaborative programme between the Dutch HSN and census projects, the main importance of the Historical Population Reports project for the

proposed VPS is that it will provide contextual statistical materials which can be utilised for research purposes alongside the VPS database. These can be used both to indicate possible areas of bias within the VPS and suggest weightings to adjust for these. It will also enable the scope of the VPS to be widened to generalised investigations of social and economic indicators in the nineteenth century.

7.2 Great Britain Historical Database

The Great Britain Historical Database (GBHD) is a large database of aggregate statistical information mainly covering the second half of the nineteenth and the early twentieth centuries. The GBHD has been created by Humphrey Southall, David Gilbert and Ian Gregory (formerly of Queen Mary College, University and now at Portsmouth University) as part of a wider Historical GIS project (see below).⁵⁹

Essentially the GBHD consists mainly of a series of computerised tables, linked by spatial units, of official statistics in Parliamentary Papers and other associated reports. In outline, the main series of available data are as follows:

- statistics from the 1861 Census and the Registrar General's reports, 1851-1861;
- employment statistics from the census, 1841-1931;
- demographic statistics from the census, 1841-1931;
- mortality statistics from the Registrar General's reports, 1861-1920;
- marriage statistics from the Registrar General's reports, 1841-1870;
- trade union statistics for the Amalgamated Society of Engineers (ASE), 1851-1918;
- official poor law statistics, 1859-1915 and 1919-1939;
- small debt statistics from county courts, 1847-1913.

A subset of the GBHD exists in an online version, available via AHDS History at the UKDA, University of Essex, from which tables of aggregate data can be downloaded for specified regions and counties.⁶⁰

The main importance of the GBHD for the proposed VPS is that it will, like the Historical Population Reports Collection (see above) provide contextual statistical materials which can be utilised for research purposes alongside the VPS database.

7.3 Historic Geographical Information Systems

This project is creating a Historical Geographical Information System (GIS) for Great Britain covering the period from the late 1830s until the early 1970s, when boundary data start to become available in digital form.⁶¹ The GIS has two major components: boundary data for the changing administrative areas of Britain; a linked database of social, economic and electoral statistics (see GBHD, above). Three key groups of boundary information have or are in the process of being digitised:

⁵⁹ See <http://www.geog.port.ac.uk/gbhgis/db/> The GBHD also incorporates various data series collected by a number of other researchers.

⁶⁰ See http://hds.essex.ac.uk/scripts/GBH_online/StandardRegion.asp. Southall, H.R., Gilbert, D.R., Gregory, I., *Great Britain Historical Database Online, 1841-1939* [computer file]. Colchester, Essex: AHDS History, UK Data Archive [distributor], 23 March 2000. SN: 33305.

⁶¹ See <http://www.geog.port.ac.uk/gbhgis/gis.htm>.

- Registration Districts / Poor Law Unions 1840/1910;
- statistical reporting units of England and Wales 1840-1973;
- civil parishes of Great Britain c.17th and post 1881.

All the GIS data, once released, is available for academic use via the EDINA UKBORDERS service at the Edinburgh University Data Library.

The main importance of the availability of historical GIS for the proposed VPS is that it will allow statistical outcomes from the VPS to be mapped and visualised spatially. This is particularly important given the 'community' dimension to the 1851 census sample on which the VPS will be based.

8 Potential partners (in addition to ESRC)

8.1 Key partners

8.1.1 The National Archives

The National Archives (TNA) are firmly committed to this project. Both of its component parts, the PRO and HMC, have a long tradition of serving HE communities and individual users and of working closely with HE partners. Online resources such as PROCAT (the online catalogue of public records) and the National Register of Archives are critical research tools for many academic disciplines. Over the past decade the PRO has acted as a the support partner to a number of projects funded by the ESRC, with outputs in the form of publications or enriched catalogue data, and it has developed clear guidelines and procedures to optimise the time and effort of academics working on its premises as research associates.

Following on from consultations with HE stakeholders in 2002 and 2003, TNA now wishes to be more proactive in working with academic partners to encourage and support the creation of major and innovative research resources drawing on its unique collections – the public records. Social scientists already use these extensively, but mainly for qualitative research, and TNA wishes to unlock their undoubtedly major potential for quantitative research as well.

The historic censuses are TNA's 'crown jewels' in terms of their very high profile and use (some two-thirds of TNA's reader visits are made in order to consult them) and high level of appeal across all user sectors. Their popularity with family historians, personal interest researchers and amateur 'micro historians' is of great potential benefit to this project.

TNA would support the VPS by:

- acting as a supporting partner and giving the full weight of its endorsement to the project;
- providing research associateship status to the project researchers permitting privileged access to the original census records and to archival advice and support;
- ensuring that the project researchers are able to access TNA's online census resources (indexes, databases, images) on the same terms as TNA's own staff;
- brokering and supporting collaborative working relationships with the key family history bodies (in particular the Federation of Family History Societies and the Society of Genealogists);
- publicising the project during set-up to attract and create incentive for potential volunteers through channels such as *Ancestors Magazine*, www.nationalarchives.gov.uk and related websites, the e-mail newsletter, and columns in family history publications;
- publicising the project once completed including, if this was felt to be appropriate, hosting an academic conference at Kew.

8.1.2 The Office for National Statistics

It is clearly too early at this stage to make any concrete statements about how the proposals in the *Changing registration* will be implemented. Yet initial discussions with members of ONS indicate that they are very interested in the research potential of the VPS and would wish to support it in the fullest terms possible.

8.1.3 General Registry Office, Scotland

It has not proved possible within the short timescale of preparing this report to ascertain a definite statement from GRO(S) regarding the VPS. It is the case that a number of academic projects in Scotland have managed to secure special access to data held by GRO(S) and a number consulted over the course of preparing this report have commented on the general willingness of GRO(S) to work in co-operation with academic projects.

8.2 Associated partners

8.2.1 FreeBMD

FreeBMD is a registered charity who have been co-ordinating the mass transcription of the indices relating to civil registration in England and Wales. Project (see section 4.2, above). They have already transcribed 51.3 million entries for the period 1851 to 1901 and expect to complete the remaining 37 million entries by the end of 2004. Early discussions with FreeBMD indicate that they are willing to work with the VPS in supplying copies of all the databases they hold.

8.2.2 QinetiQ

QinetiQ produced and maintain the online version of the 1901 census of England and Wales on behalf of TNA. Initial discussion with QinetiQ has confirmed that they may be willing to provide the VPS with a version of the 1901 database for purpose of linking. However, it is important to realise that this arrangement may require a payment to QinetiQ.

8.3 Volunteer groups

Initial meetings with representatives of the Federation of Family History Societies, Society of Genealogists and the British Association for Local History have been encouraging. All seems keen to support the project by promoting it and helping to find volunteer researchers.

8.4 Potential AHRB involvement

Discussions with the AHRB suggest that they may be interested in becoming a funding partner in the VPS. If the AHRB were to be involved then clearly a case would need to be made that the VPS would be valuable for humanities as well as social science research. In particular, the AHRB are interested in the possibilities for research created by the community feature of the 1851 census sample, and the potential for building community profiles overtime. This could include data enrichment and enlargement beyond the core dataset, for the analysis of communities, real space, religion and religious congregations, associations, ethnicity, and much else, including the repopulating of houses and spaces. Thus, if a successful case could be made, the AHRB may be willing to support the creation of intercensal sub-sets which would enable more intensive enrichment, allowing for historical and cultural geographical work to be carried out alongside the main body of the VPS.

8.5 International collaborations

8.5.1 Minnesota Population Center

The Director of the Minnesota Population Center, Professor Steven Ruggles, has already agreed to work openly with the VPS. They will make available any automated record linkage software that they develop, and provide access to staff with expertise in related areas. Should the VPS proceed, Ruggles has already indicated that he would be interested in applying to the NFS to support a joint research programme.

8.5.2 Historical Sample of the Population of the Netherlands

The HSN has indicated that they would welcome working with the VPS. In particular they would share their experience on sampling, linking and refreshment and make available data collection software and database designs for adaptation as appropriate. They are also keen to share information on coding and classification schemes in order to facilitate research across the two data collections.

8.6 E-Science

Although it is difficult at this stage to predict now the VPS might develop, should it be funded, it is important to note that the project could potentially open up significant opportunities for e-Science related research activity. First, the sheer size of the project, with record linkage being undertaken on a database of up to 257 million records, suggests that grid technologies may be important for computational processing. Second, it may prove beneficial to create interoperability between the VPS and other existing data resources, especially if the AHRB fund the creation of community profile sub-sets. Such resources are likely to be disparate and would potentially benefit from being grid enabled.

9 Recommendations

This report, out of necessity given the time constraints, has identified various possibilities, investigated various problems, posed various questions, but has not provided many answers. In order to investigate further some of the methodological issues raised in this report and to experiment working with a volunteer workforce it is proposed that the next step in assessing the possibility of a VPS should be to undertake a pilot project.⁶² A successful pilot project will be able to reveal potential problems before an extension to national coverage is instigated, assuming that such a decision is taken.

The proposed pilot project would, in effect, attempt to undertake a VPS in miniature, but in so doing would concentrate on addressing problems and testing different methodological approaches rather than trying to create a fully representative small scale VPS. In particular the pilot project should:

- discussion possibilities further with potential stake holders;
- investigate potential user needs;
- develop working arrangements with volunteer researchers;
- examine different cohort and panel approaches;
- experiment with different 'target' and 'observation' strategies;
- explore the problems of refreshment;
- prepare initial designs for a linkage database;
- investigate possible automated record linkage strategies;
- assess coding and standardisation schemes used in other longitudinal studies;
- investigate what other source materials could feasibly be linked to a VPS subset.

In terms of data it is further proposed that the pilot study concentrate on two geographical areas: Oxfordshire and Norfolk. The principal reason for this choice is that important computerised sources already exist for these two counties. Oxfordshire already has all of the censuses of 1851-1891 indexed as well as all the civil registration information for this period. This is the only county in England and Wales with such a complete coverage. Norfolk is suggested because a large amount of CEB information already exists in database form, which could be useful to test linkage routines. Oxfordshire, in particular also has a very active family history society from which it would be hoped to recruit and test volunteers.

The details of the EDs for these two counties captured in the 1851 sample are given below. Although neither county is particularly industrial or urban, in combination the two counties provide a wide range of community types from urban Norwich (the country's 10th largest city in 1851), the port of King's Lynn, the academic seat of Oxford, as well as a mix of large to small villages.

It is proposed that in order to meet the objectives set out above a pilot project might run for 24 months, employing two RAs and a part-time project manager, at an estimated cost of £220-250k.

⁶² The importance of a constructive pilot project has also been highlighted in the case of the HSN.

Norfolk

- Grimstone,
- Hethersett,
- Hickling,
- Hockwold Cum Wilton,
- Kerdiston,
- King's Lynn - All Saints,
- King's Lynn - South Lynn,
- Norwich - St Mary Coslany,
- Norwich - St Stephen,
- Ranworth,
- Seething,
- Southburgh,
- Stiffkey,
- Tacolneston,
- Taverham,
- Tilney St Lawrence,
- Wereham,
- Whissonsett

Stratification type	Population	n. of households
'Communities' under 2,000 population	8,188	1,784
'Towns' and municipal Boroughs	2,477	612
Institutions	102	*11
Total	10,767	**2,396
* This figure is the total number of institutions in the Norfolk sample		
** This figure excludes the 11 institutions		

Oxfordshire

- Ambrosden,
- Churchill,
- Combe,
- Horton Cum Studley,
- Oxford - St Mary Magdalene,
- Rofford,
- Shotswell,
- Weald

Stratification type	Population	n. of households
'Communities' under 2,000 population	2,420	530
'Towns' and municipal Boroughs	780	144
Parliamentary Boroughs	653	138
Institutions	78	*6
Total	3,931	**812
* This figure is the total number of institutions in the Oxfordshire sample		
** This figure excludes the 6 institutions		

10 Bibliography

10.1 Published works

Adman, P., Baskerville, S. W. and Beedham, K. F. 'Computer-assisted record linkage: or how best to optimize links without generating errors' *History & Computing*, **4** (1992) 2-15.

Anderson, M. 'Marriage patterns in Victorian Britain: an analysis based on registration district data for England and Wales, 1861', *Journal of Family History*, **1** (1976) 55-79.

Anderson, M. *National sample from the 1851 census of Great Britain: introductory user guide*, (University of Edinburgh, Dept. of Economic and Social History, 1987)

Anderson, M., Kemmer, D. and Morse, D. J. *Demographic change in Scotland 1855-1914: the onset of the fertility decline in Scotland, some results from an exercise in family reconstruction*, Working paper 2, (version 2), Department of Economic and Social History and University of Edinburgh Data Library, (University of Edinburgh, 1992).

Annal, D. *Using birth, marriage and death records*, (Richmond, 2002).

Baines, D. *Migration in a mature economy: emigration and internal migration in England and Wales, 1861-1900*, (Cambridge, 1985).

Bloothooft, G. 'Assessment of systems for nominal retrieval and historical record linkage', *Computers and the Humanities*, **32** (1998) 39-56.

Bouchard, G. 'Current issues and new prospects for computerized record linkage in the Province of Quebec', *Historical Methods*, **25** (1992) 67-73.

Bouchard, G. 'The Processing of Ambiguous Links in Computerized Family Reconstruction', *Historical Methods*, **19** (1986) 9-19.

Bynner, J., Butler, N., Ferrie, E., Shepherd, P. and Smith, K. *The design and conduct of the 1999-2000 survey of the National Child Development Study and the 1970 British Cohort Study*, Centre for Longitudinal Studies, Working Paper no. 1 (London, n.d.).

Crawford, E. M. *Counting the people. A survey of the Irish censuses, 1813-1911*, (Dublin, 2003).

Crompton, C. A. 'An Exploration of the Craft and Trade Structure of Two Hertfordshire Villages, 1851-1891: an Application of Nominal Record Linkage to Directories and Census Enumerators Books', *The Local Historian*, **28** (1998) 145-58.

De Brou, D. and Olsen, M. 'The Guth Algorithm and the Nominal Record Linkage of Multi-Ethnic Populations', *Historical Methods*, **19** (1986) 20-4.

Dennis, R. J. 'Distance and social interaction in a Victorian city', *Journal of Historical Geography*, **3** (1977) 237-50.

Dennis, R. J. 'Intercensal mobility in a Victorian city', *Transactions of the Institute of British Geographers*, new series **2** (1977) 349-63.

Despotidon, S. and Shepherd, P. 1970 *British Cohort Study twenty-six year follow-up. Guide to data available at the ESRC Data Archive*, (London, n.d.).

Dixon, B. *Birth and death certificates. England and Wales 1837 to 1969*, (Burnham, 1999).

Dixon, B. *Marriages and certificates in England & Wales*, (Burnham, 2000).

Ferrie, E. ed *Life at 33: the fifth follow-up of the National Child Development Study*, (London, 1993).

Ferrie, J. 'A new sample of males linked from the Public-Use-Microdata-Sample of the 1850 US Federal Census of Population to the 1860 US Federal Census Manuscript Schedules', *Historical Methods*, **29** (1996) 141-56.

Ferrie, J. *How ya gonna keep 'em down on the farm [when they've seen Schenectady]?* Rural to urban migration in nineteenth-century America 1850-1870. Working paper, (Northwestern University, 1999).

Fogelvik, S. 'The Stockholm Historical Database at work', in Denley, P., Fogelvik, S. and Harvey, C. eds, *History and Computing II*, (Manchester, 1989) 256-65.

Friedlander, D. and Roshier, R. J. 'A Study of Internal Migration in England and Wales: Part 1', *Population Studies*, **19** (1966) 239-79.

Garrett, E. 'Trials of labour: motherhood versus employment in a nineteenth-century textile centre', *Continuity and Change*, **5** (1990) 121-54.

Guth, G. J. A. 'Surname Spelling and Computerized Record linkage', *Historical Methods Newsletter*, **10** (1976) 10-9.

Hallas, C. S. 'The social and economic impact of a rural railway: the Wensleydale line', *Agricultural History Review*, **34** (1986) 29-44.

Hancock, R. 'In Service or One of the Family? Kin-Servants in Swavesey 1851-1881, Ryde 1881, and Stourbridge 1881', *Family and Community History*, **2** (1999) 141-18.

Harvey, C. and Green, E. 'Record Linkage Algorithms: Efficiency, Selection and Relative Confidence', *History and Computing*, **6** (1994) 143-52.

Harvey, C., Green, E. M. and Corfield, P. J. 'Record Linkage Theory and Practice: an Experiment in the Application of Multiple Pass Linkage Algorithms' *History and Computing*, **8** (1996) 78-89.

Hattersley, L. and Creeser, R. *Longitudinal study 1971-1991. History, organisation and quality of data*, OPCS Series LS no. 7 (London, 1995).

Hershberg, T., Burstien, A. and Dockhorn, R. 'Record Linkage', *Historical Methods Newsletter*, **2/3** (1976) 137-63.

Higgs, E. *A clearer sense of the census. The Victorian censuses and historical research*, (London, 1996).

Higgs, E. *Making sense of the census: the manuscript returns of the census 1801-1901*, (London, 1989).

Hinde, P. R. A. 'The population of a Wiltshire village in the nineteenth century: a reconstruction study of Berwick St James, 1841-1871', *Annals of Human Biology*, **14** (1987) 475-85.

Katz, M. and Tiller, J. 'Record linkage for everyman: a semi-automatic process', *Historical Methods Newsletter*, **5** (1972) 144-50.

Kitts, A., Doulton, D. and Reis, E. *The reconstruction of Viana do Castelo* (Egham, 1990).

Lawton, R. 'Population changes in England and Wales in the later nineteenth century: an analysis of trends by registration districts', *Transactions of the Institute of British Geographers*, **44** (1968) 55-74.

Lawton, R. ed. *The census and social structure: an interpretative guide to 19th century censuses for England and Wales*, (London, 1978).

Lawton, R. 'Census data for urban areas', in Lawton, R. ed. *The census and social structure*, 82-145.

Mandemakers, K. 'Historical Sample of the Netherlands (HSN). Background, Objectives and International Context.', in Marker, H. J. and Pagh, K. eds, *Yesterday. Proceedings from the 6th International Conference Association of History and Computing, Odense 1991*, (Odense, 1994) 174-81.

Mandemakers, K. 'The Netherlands. Historical Sample of the Netherlands', in Kelly Hall, McCaa, R. and Thorvaldsen, G. eds, *Handbook of International Historical Microdata for Population Research*, (Minneapolis, 2000) 149-77.

Mandemakers, K. 'The Historical Sample of the Netherlands' *Historical Social Research*, **26** (2001) 179-90.

Miller, R. and Thorvaldsen, G. 'Beyond Record Linkage: Longitudinal Analysis of Turn-of-the-century Inter-urban Swedish Migrants', *History and Computing*, **9** (1997) 106-21.

Mills, D. and Schürer, K. eds *Local communities in the Victorian census enumerators' books*, (Oxford, 1996).

Mills, D. R. and Pearce, C. *People and places in the Victorian census. A review and bibliography of publications based substantially on the manuscript census enumerators' books, 1841-1911*, Institute of British Geographers, Historical Geography Research Series, No 23, (Cheltenham, 1989).

Mineau, G. P., Bean, L. L. and Anderton, D. L. 'Description and evaluation of linkage of the 1880 census to family genealogies: implications for Utah fertility research', *Historical Methods*, **22** (1989) 144-57.

Morton, G. 'Presenting the self: record linkage and referring to ordinary historical persons', *History and Computing*, **6** (1994) 12-20.

Nenadic, S. 'Studying the middle class in nineteenth-century urban Britain', *Urban History Yearbook*, (1987) 22-41.

- Nissel, M. *People count: a history of the General Register Office*, (London, 1987)
- Office for National Statistics, *Civil registration: delivering vital change*, (London, 2003).
- Perkyns, A. 'Age checkability and accuracy in the censuses of six. Kentish parishes, 1851-1881', in Mills and Schürer eds *Local communities*, 115-34.
- Perkyns, A. 'Birthplace accuracy in the censuses of six. Kentish parishes, 1851-1881', in Mills and Schürer eds *Local communities*, 229-45.
- Pooley, C. G. 'Residential mobility in the Victorian city', *Transactions of the Institute of British Geographers*, new series **4** (1979) 258-77.
- Pooley, C. G. and Doherty, J. C. 'The longitudinal study of migration. Welsh migration to English towns in the nineteenth century', in Pooley and Whyte eds *Migrants, emigrants and immigrants*, 143-73.
- Pooley, C. G. and Turnball, J. *Migration and mobility in Britain since the eighteenth century*, (London, 1998).
- Pooley, C. G. and Whyte, I. D. eds *Migrants, emigrants and immigrants. A social history of migration*, (London, 1991).
- Reay, B. 'Before the transition: fertility in English villages, 1800-1880', *Continuity and Change*, **9** (1994) 91-120.
- Reay, B. 'Kinship and the neighbourhood in nineteenth-century rural England: the myth of the autonomous nuclear family', *Journal of Family History*, **21** (1996) 87-104.
- Royle, S. A. 'Irish manuscript census records: a neglected source of information', *Irish Geography*, **2** (1978) 110-25.
- Russell, N. and Phelps, A. *Youth Cohort Study. Cohort 10 Sweep 1 (C10S1). Technical report* (2001).
- Schürer, K. 'The role of the family in the process of migration', in Pooley and Whyte eds, *Migrants, emigrants and immigrants*, 106-42.
- Shepherd, P., Smith, K., Joshi, H. and Dex, S. *Millennium Cohort Study first survey: a guide to the SPSS dataset*, Centre for Longitudinal Studies (London, 2003).
- Smith, K. and Joshi, H. 'The Millennium Cohort Study', *Population Trends*, **107** (2002) 30-4.
- Spencer, A. 'Scotlands people web site', *Computers in Genealogy*, **8** (2003), 55-7.
- Stenflo, G. and Sundin, J. 'Using a large historical database. An example from the Demographic Database in Umeå', in Denley, P. and Hopkin, D. eds, *History and Computing*, (Manchester, 1987) 58-62.
- Taylor, I. C. 'Liverpool's institutional and quasi-institutional populations in 1841 and 1851', in Mills and Schürer eds, *Local communities*, 42-6.
- Taylor, M. F. ed. with Brice, J., Buck, N. and Prentice-Lane, E. *British Household Panel Study user manual. Volume A. Introduction, technical report and appendices*, (Colchester, 2002).

- Tetielbaum, M. S. *The British fertility decline*, (Princeton, 1984).
- Tilley, P. and French, C. 'Record Linkage for Nineteenth-Century Census Returns: Automatic or Computer-Aided?' *History and Computing*, **9** (1997) 123-33.
- White, M. B. 'Family migration in Victorian Britain: the case of Grantham and Scunthorpe', in Mills and Schürer eds, *Local communities*, 267-77.
- Wojciechowska, B. 'Brenchley: a study of migratory movements in a mid-nineteenth-century rural parish', in Mills and Schürer eds, *Local communities*, 253-66.
- Woods, R. I. 'Approaches to the fertility transition in Victorian England', *Population Studies*, **41** (1987) 283-311.
- Woods, R. I. *The demography of Victorian England and Wales*, (Cambridge, 2000).
- Wrigley, E. A. ed. *Identifying people in the past* (London, 1973).
- Wrigley, E. A., Davies, R. S., Oeppen, J. O. and Schofield, R. S. *English population history from family reconstitution, 1580-1837*, (Cambridge, 1997).
- Wrigley, E. A. and Schofield, R. S. *The population history of England 1541-1871: a reconstruction*, (Cambridge, 1981).

10.2 Unpublished reports

- Bengtsson, T. and Lundh, C. *Evaluation of a computer program for automatic family reconstruction*, Working paper presented at a seminar of the Cambridge Group for the History of Population and Social Structure, 6 May 1991.
- Drake, M. 'Ashford 1840-1870: a socio-demography study', (unpublished Final Report, Centre for Research in the Social Sciences in the University of Kent at Canterbury, 1970).
- Lait, A.J., and B. Randell, 'An assessment of name matching algorithms', <http://homepages.cs.nc1.ac.uk/brian.Randell/home.informal/Genealogy/NameMatching.pdf> , 14-08-03.

10.3 Web sites consulted

- <http://www.rhd.uit.no/nhdc/micro.html>
Historical micro-data from around the world.
- <http://www.iisg.nl/~hsn/>
The Historical Sample of the Netherlands:
- <http://www.ddb.umu.se/index.html>
The Demographic Database at Umeå:
- <http://www.genealogie.umontreal.ca/fr/main.htm>
PRDH

<http://www.cls.ioe.ac.uk/Ls/lshomepage.html>

The Longitudinal Study

<http://www.fcsn.gov/working-papers/hattersley.pdf>

The Longitudinal Study: Creation of the sample and methods of linkage.

<http://www.iser.essex.ac.uk/bhps/index.php>

BHPS

<http://www.iser.essex.ac.uk/bhps/doc/vola/app5.html>

BHPS: Related publications and documentation.

<http://www.cls.ioe.ac.uk/Ncds/nhome.htm>

The National Child Development Survey

<http://www.statistics.gov.uk/STATBASE/Product.asp?vlnk=1213>

Youth Cohort Study

<http://www.dfes.gov.uk/cgi-bin/rsgateway/search.pl?keyw=078&q2=Search>

Youth Cohort Study: Reports.

<http://www.cls.ioe.ac.uk/Bcs70/bhome.htm>

The 1970 British Cohort Study

http://www.mimas.ac.uk/system/news/BCS70_DOCS.html

The 1970 British Cohort Study: Documentation.

http://hds.essex.ac.uk/scripts/GBH_online/StandardRegion.asp

Great Britain Historical Database Online

<http://www.geog.port.ac.uk/gbhgis/>

Great Britain Historical GIS Programme

<http://www.scotlandspeople.gov.uk/>

GRO Scotland home page.

<http://www.gro-scotland.gov.uk/grosweb/grosweb.nsf/pages/cen41-71>

Coverage of indexed censuses.

<http://www.scotlandsclans.com/genscot.htm>

ScotsClans.com. search.

<http://www.scotsorigins.com/>

Scots Origins search.

<http://www.scotlandsclans.com/census.htm>

Ancestry.com. search.

<http://www.ofhs.org.uk/>

Oxfordshire Family History Society (census indexes).

<http://www.jenlibrary.u-net.com/>

The Family History Shop.

<http://www.genealogysupplies.com/>

S&N Genealogy Supplies.

<http://www.stepping-stones.co.uk/>

Stepping Stones.

<http://www.britishdataarchive.com/>

British Data Archive.

<http://www.thegenealogist.co.uk/>

The Genealogist: UK census on line.

<http://freebmd.rootsweb.com/>

FreeBMD.

<http://citeseer.nj.nec.com/context/2114780/0>

Outlines of theoretical articles on record linkage.

10.4 Data collections

Anderson, M. et al. *National Sample from the 1851 Census of Great Britain* [computer file]. Colchester, Essex: UK Data Archive [distributor], 1979. SN: 1316.

Bynner, J.M. *1970 British Cohort Study: Twenty Six-Year Follow-Up, 1996* [computer file]. 2nd ed. Colchester, Essex: UK Data Archive [distributor], 22 January 2003. SN: 3833.

City University Social Statistics Research Unit, *National Child Development Study Composite File Including Selected Perinatal Data and Sweeps One to Five, 1958-1991* [computer file]. 2nd ed. National Birthday Trust Fund, National Children's Bureau, City University Social Statistics Research Unit [original data producers]. Colchester, Essex: UK Data Archive [distributor], 25 July 2000. SN: 3148.

Economic and Social Research Council Research Centre on Micro-Social Change, *British Household Panel Survey* [computer file]. Colchester, Essex: UK Data Archive [distributor], 6 May 2003. SN: 4651.

Joint Centre for Longitudinal Research, *National Child Development Study and 1970 British Cohort Study (BCS70) Follow-ups, 1999-2000* [computer file]. 2nd ed. Colchester, Essex: UK Data Archive [distributor], 29 January 2003. SN: 4396.

Pooley, C.G. and Turnbull, J., *Longitudinal Study of Residential Histories, 1750-1994* [computer file]. Colchester, Essex: UK Data Archive [distributor], 30 July 1997. SN: 3571.

Schürer, K. and Woollard, M. *1881 Census for England and Wales, the Channel Islands and the Isle of Man (Enriched Version)* [computer file]. Genealogical Society of Utah, Federation of Family History Societies [original data producers]. Colchester, Essex: UK Data Archive [distributor], 2000. SN:4177

Schürer, K. and Woollard, M. *1881 Census for Scotland (Enriched Version)* [computer file]. Genealogical Society of Utah, Federation of Family History Societies [original data producers]. Colchester, Essex: UK Data Archive [distributor], 2000. SN:4178.

Southall, H.R., Gilbert, D.R. and Gregory, I. *Great Britain Historical Database Online, 1841-1939* [computer file]. Colchester, Essex: AHDS History, UK Data Archive [distributor], 23 March 2000. SN: 33305.

11 Individuals and groups consulted

The following lists the individuals and groups consulted during the course of preparing this report.

Professor M. Anderson	University of Edinburgh
Professor N. Buck	ULSC, University of Essex
Professor G. Crossick	Arts and Humanities Research Board
Professor A. Dale	Cathy Marsh Centre for Census and Survey Research, Manchester.
Drs P. Doorn	Netherlands Institute for Scientific Information Services (NIWI)
Dr E. Garrett	Universities of Cambridge and Portsmouth
Dr E. Hallam Smith	The National Archives
Professor P. Lynn	ISER, University of Essex
Drs C. Mandemakers	International Institute of Social History, Netherlands
Professor C. Pooley	University of Lancaster
Mr John Pullinger	Office for National Statistics
Professor S. Ruggles	University of Minnesota
Matthew Woollard	AHDS History, University of Essex
Professor Sir E. A. Wrigley	University of Cambridge

British Association for Local History
Federation of Family History Societies
Society of Genealogists